

# Implications of uniformly distributed, empirically informed priors for phylogeographical model selection: A reply to Hickerson et al.

Jamie R. Oaks<sup>\*†1</sup>, Charles W. Linkem<sup>2</sup>, and Jeet Sukumaran<sup>3</sup>

<sup>1</sup>Department of Ecology and Evolutionary Biology, University of Kansas,  
Lawrence, Kansas 66045

<sup>2</sup>Department of Biology, University of Washington, Seattle, Washington 98195

<sup>3</sup>Department of Ecology and Evolutionary Biology, University of Michigan,  
Ann Arbor, Michigan 48109

February 27, 2014

**Running head:** Approximate Bayesian model choice

## Abstract

Biogeographers frequently seek to explain population and species differentiation on geographical phenomena. Establishing that a set of population splitting events occurred at the same time can be a potentially persuasive argument that a set of taxa were affected by the same geographic events. Unfortunately, estimating divergence times precisely when one lacks precise information about the rate of molecular evolution is a notoriously difficult problem in evolutionary biology. Huang et al. (2011) introduced an approximate Bayesian approach (implemented in the software **msBayes**) which can be used to estimate the probabilities of models in which multiple sets of taxa diverge at the same time. Recently, Oaks et al. (2013) used this model-choice framework to study 22 pairs of vertebrate lineages which are distributed across the Philippines; they also studied the behavior of the **msBayes** approach using computer simulations. Oaks et al. (2013) found that the model was very sensitive to the prior and that the inference method had low power to detect variation in divergences times. These results were not surprising, in light of a rich statistical literature showing that the marginal likelihood of a model is sensitive to the use of vague priors. Because this sensitivity to prior assumptions affects the crucial insights that a researcher who employs **msBayes** seeks to gain, Oaks et al. (2013) recommended that users of the approach should carefully assess the robustness of their conclusions to different priors. According to Hickerson et al.

---

<sup>\*</sup>Corresponding author: joaks1@gmail.com

<sup>†</sup>Current address: Department of Biology, University of Washington, Seattle, Washington 98195

(2014), the lack of robustness in **msBayes** analyses was due to excessively broad priors on divergence times, leading to inadequate numbers of simulation replicates. They proposed a new model-averaging approach that uses narrow, empirically informed uniform priors. Here we demonstrate that the approach of Hickerson et al. (2014) is dangerous in the sense that the empirically-derived priors often exclude from consideration the true values of the models' parameters. On a more fundamental level, we question the value of adopting an empirical Bayesian stance for this model-choice problem, because it can mislead model posterior probabilities, which are inherently measures of belief in the models after prior knowledge is updated by the data. The robust Bayesian approach of conducting analyses under a variety of priors can reveal prior sensitivity and communicate which assumptions underlie model inference. Furthermore, simulations provide insight into the temporal resolution of the method, which in turn helps guide interpretation of results.

**KEY WORDS:** Approximate Bayesian computation; Bayesian model choice; empirical Bayes

# 1 Introduction

**msBayes** (Huang et al., 2011) uses approximate Bayesian computation (ABC) to estimate the distribution of divergence times among co-distributed pairs of taxa. It approximates a posterior probability over models that range from a single divergence-time parameter (i.e., simultaneous divergence of all pairs of taxa) to the fully generalized model in which each pair of taxa diverged at a unique time. A full description of the model is given in Oaks et al. (2013). Frequently, the exact values of the divergence times are hard to determine because researchers typically lack precise knowledge of the rate at which substitutions occur. The exact divergence time is therefore often viewed as a nuisance parameter, and researchers focus on the number of divergence events. The papers of Oaks et al. (2013) and Hickerson et al. (2014) are in agreement on the fundamental methodological point about the model selection performed in **msBayes**:

- The use of vague priors on the divergence-time parameters can lead to support for models with few divergence events shared across taxa. Thus, the primary inference enabled by the approach is very sensitive to the priors on divergence times.

This is not surprising given the rich statistical literature that shows that marginal likelihoods are very sensitive to the priors used in Bayesian model selection (e.g., Jeffreys, 1939; Lindley, 1957). Accordingly, Oaks et al. (2013) suggest the primary cause of spurious support for models with few divergence parameters is the greater marginal likelihoods of these models under vague uniform priors, relative to more parameter-rich models. Models with more divergence-time parameters integrate over *much* greater parameter space with low probability of producing the data, yet relatively high prior density imposed by the uniformly distributed prior on divergence times. Note, this is a statistical issue, extrinsic to **msBayes**, and does not question the soundness of the numerical approximation machinery of **msBayes**.

Hickerson et al. (2014) suggest the bias is the result of a numerical issue intrinsic to **msBayes**, concluding the method’s rejection algorithm is inefficient and will be biased toward models with fewer divergence-time parameters, because the parameter space of these models are more densely sampled relative to models with more divergence-time parameters. We agree that sampling error is present in all numerical Bayesian estimation methods, however, we show here that it is not a major contributor to the biases found by Oaks et al. (2013). We present results here of analyses that, when combined with the results of Oaks et al. (2013), strongly suggest that **msBayes** prefers models with few divergence-time parameters, because they have greater likelihoods when averaged over broad uniform priors on divergences times.

Oaks et al. (2013) suggest alternative prior probability distributions on divergence-times and other nuisance parameters could increase the marginal likelihoods of models with more divergence-time parameters and thus reduce spurious support for models of temporally clustered divergences. Alternatively, Hickerson et al. (2014) present an approach that uses empirically informed uniform priors and Bayesian model-averaging in an attempt to accommodate uncertainty in selecting priors. In general, we agree with the use of Bayesian model averaging to obtain a posterior estimate marginalized over uncertainty in prior choice. However, we question whether adding an additional dimension of model choice that samples over empirically informed uniform priors is a fruitful solution for a model-choice method that is highly sensitive to priors.

In this paper, we discuss the potential theoretical and practical considerations of using empirically informed priors for Bayesian model choice. Furthermore, we evaluate the empirical Bayesian model-averaging approach of Hickerson et al. (2014) as a potential solution to the biases of `msBayes` revealed by Oaks et al. (2013). In their re-analysis of the dataset of Oaks et al. (2013), Hickerson et al. (2014) made an error by mixing different units of time, which makes the results presented in their response difficult to interpret (see Supporting Information for details). Here, we avoid this error, but still find their approach to be biased toward clustered divergence models. Furthermore, the approach provides another means by which the method can “escape” models with large parameter space, which manifests in the preference of models that exclude from consideration the true values of the model’s parameters. Our results, when combined with those of Oaks et al. (2013), suggest that it is difficult to choose a uniformly distributed prior on divergence times that is broad enough to confidently contain the true values of parameters while being narrow enough to avoid spurious support of models with less parameter space. However, recent work shows more flexible probability distributions without a hard upper bound (e.g., gamma) can accommodate prior uncertainty in divergence times without inhibiting the marginal likelihoods of models with more divergence-time parameters, and as a result, increase the method’s power to detect temporal variation in divergences Oaks (2014).

## 2 The potential implications of empirical Bayesian model choice

Hickerson et al. (2014) repeatedly refer to the prior distributions used by Oaks et al. (2013) for divergence-time parameters as “poorly selected.” However, this is misleading, as Oaks et al. (2013) discuss in detail how they selected their prior to reflect the large amount of uncertainty about the timing of divergences across all 22 of the taxa in their study (see section “Specifying and simulating the joint prior” of Oaks et al. (2013)). Oaks et al. (2013) further discuss how the use of uniform prior distributions in `msBayes` for many of the model’s parameters requires investigators to select broad priors to avoid excluding the truth *a priori*, resulting in too much density in improbable regions of parameter space.

Hickerson et al. (2014) suggest that Oaks et al. (2013) should have used their data to inform their prior on divergence times (i.e., an empirical Bayesian approach). However, Oaks et al. (2013) did use empirically informed priors to mimic empirical situations in which a large amount of prior information is available; they did so to assess the prior sensitivity of the method and to determine if uniform priors in such situations are narrow enough to avoid spurious support of models with few divergence-time parameters. The results of Oaks et al. (2013) and Hickerson et al. (2014) show the method is highly sensitive to the prior on divergence times. Furthermore, Oaks et al. (2013) found the method remained biased toward clustered divergences under the empirically informed priors.

The main argument of Hickerson et al. (2014) is that the priors used by Oaks et al. (2013) were not informative enough. They suggest a very narrow, highly informed uniform prior on divergence times is necessary to avoid the method’s preference for models with few divergence-time parameters. Such an empirical Bayesian approach to model selection raises some theoretical and practical concerns, some of which were discussed by Oaks et al. (2013)

(see the last paragraph of “Assessing prior sensitivity of msBayes” in Oaks et al. (2013)); we expand on this here.

## 2.1 Theoretical implications of empirical priors for Bayesian model choice

Bayesian inference is a method of inductive learning in which Bayes’ rule is used to update our beliefs about a model  $M$  as new information becomes available. If we let  $\Theta$  represent the set of all possible parameter values for model  $M$ , we can define a prior distribution for all  $\theta \in \Theta$  such that  $p(\theta | M)$  describes our belief that any given  $\theta$  is the true value of the parameter. If we let  $\mathcal{X}$  represent all possible datasets then we can define a sampling model for all  $\theta \in \Theta$  and  $X \in \mathcal{X}$  such that  $p(X|\theta, M)$  measures our belief that any dataset  $X$  will be generated by any state  $\theta$  of model  $M$ . After collecting a new dataset  $X_i$ , we can use Bayes’ rule to calculate the posterior distribution

$$p(\theta | X_i, M) = \frac{p(X_i | \theta, M)p(\theta | M)}{p(X_i | M)}, \quad (1)$$

as a measure of our beliefs after seeing the new information, where

$$p(X_i | M) = \int_{\theta} p(X_i | \theta, M)p(\theta | M)d\theta \quad (2)$$

is the marginal likelihood of the model.

This is an elegant method of updating our beliefs as data are accumulated. However, this all hinges on the fact that the prior ( $p(\theta | M)$ ) is defined for all possible parameter values independently of the new data being analyzed. Any other datasets or external information can safely be used to inform our beliefs about  $p(\theta | M)$ . However, if the same data are used to both inform the prior and calculate the posterior, the prior becomes conditional on the data, and Bayes’ rule breaks down.

Thus, empirical Bayes methods have an uncertain theoretical basis and do not yield a valid posterior distribution from Bayes’ rule (e.g., empirical Bayesian estimates of the posterior are often too narrow, off-center, and incorrectly shaped; Morris, 1983; Laird and Louis, 1987; Carlin and Gelfand, 1990; Efron, 2013). This is not to say that empirical Bayesian approaches are not useful. Empirical Bayes is a well-studied branch of Bayesian statistics that has given rise to many inference methods that provide means of parameter estimation that often exhibit favorable frequentist properties. Furthermore, many post-hoc correction methods have been developed for estimating confidence-intervals from empirical Bayes estimates of distributions that often exhibit well-behaved frequentist coverage probabilities (Morris, 1983; Laird and Louis, 1987, 1989; Carlin and Gelfand, 1990; Hwang et al., 2009).

Whereas empirical Bayes approaches can provide powerful methods for parameter estimation, a theoretical justification for empirical Bayes approaches to model choice is questionable. In Bayesian model choice, the primary goal is not to estimate parameters, but to estimate the relative probabilities of candidate models. In a simple example where two candidate models,  $M_1$  and  $M_2$ , are being compared, the goal is to estimate the posterior

probabilities of these two models. Again, we can use Bayes' rule to calculate this as

$$p(M_1 | X_i) = \frac{p(X_i | M_1)p(M_1)}{p(X_i | M_1)p(M_1) + p(X_i | M_2)p(M_2)}. \quad (3)$$

By comparing Equations 1 and 3, we see fundamental differences between Bayesian parameter estimation and model choice. In Equation 1, we see that the posterior density of any state  $\theta$  of the model, is the prior density updated by the probability of the data given the state  $\theta$  (the likelihood of  $\theta$ ). The integral over the entire parameter space of the model, which is defined by the priors, only appears as a normalizing constant in the denominator. Thus, as long as the prior distribution contains the values of  $\theta$  that maximize the probability of the data (i.e., the maximum likelihood) and the data are strongly informative relative to the prior, the values of the parameters that maximize the posterior distribution will be relatively robust to prior choice, even if the posterior distribution is technically incorrect due to using the data to inform the priors. This is why empirical Bayes can work well for estimating the values of model parameters.

However, if we look at Equation 3, we see that in Bayesian model choice it is now the *marginal* likelihood of a model that updates the prior to yield the model's posterior probability. Thus the integral over the entire parameter space of the likelihood weighted by the prior probability density is no longer a normalizing constant, but rather is what informs the posterior probability of that model from the data. As a result, Bayesian model choice tends to be much more sensitive to priors than parameter estimation.

Another important difference of Bayesian model choice illustrated by Equation 3 is that the value of interest, the posterior probability of a model, is not a function of  $\theta$ , because it is integrated out of the marginal likelihoods of the candidate models. Thus, unlike parameter estimates, the estimated posterior probability of a model is a single value (rather than a distribution) lacking a measure of posterior uncertainty. Also, unlike model parameters, the posterior probabilities of candidate models have no clear true values. Model posterior probabilities are inherently measures of our belief in the models after our prior beliefs are updated by the data being analyzed.

This complicates the meaning of model posterior probabilities when Bayes' rule is violated by informing priors with the same data to be analyzed. For empirical Bayesian parameter estimation, the objective is that by giving the data more weight relative to the prior, the posterior distribution will be peaked near the true value(s) of the model's parameter(s). There is no such justification for empirical Bayesian model choice, because there are no true values for the model probabilities being estimated. By using the data twice, we fail to account for prior uncertainty and mislead our posterior beliefs in the models being compared; we will be over confident in some models and under confident in others. Nonetheless, empirical Bayesian model choice does perform well for some problems. Particularly, in cases where large aggregate data sets are used for many parallel model-choice problems simultaneously, pooling information to inform priors can lead to favorable group-wise frequentist coverage across tests (Efron, 2008). However, this is far removed from the single model-choice problem of `msBayes`.

These distinctions between Bayesian parameter estimation and model choice can be illustrated with a simple example. Let us say we are interested in the fairness of a particular coin, and we denote the unknown probability of it landing heads as  $\theta$ . More specifically, we

are interested in the probability of two models,  $M_1$  and  $M_2$ . In both models the outcomes of flipping the coin are assumed to be binomially distributed, but under  $M_1$  the coin is weighted toward landing heads (i.e.,  $\theta > 0.5$ ), whereas under  $M_2$ , the coin is weighted toward landing tails (i.e.,  $\theta < 0.5$ ). We already have data from flipping a different coin 20 times that landed both heads and tails 10 times each, and so we decide to use these data in specifying a beta prior on fairness of the new coin of  $beta(a = 10, b = 10)$  (Figure 1). We collect data by flipping the coin of interest  $N = 10$  times,  $y = 3$  of which land heads. Given the beta distribution is a conjugate prior for a binomial likelihood, the posterior distribution has the nice analytical form  $\theta | y, N \sim beta(a + y, b + N - y)$ , which for the new dataset is simply  $beta(13, 17)$  (Figure 1). The maximum a posteriori (MAP) estimate of the probability of heads is 0.429, and following Equation 2 the marginal likelihoods of our models of interest are

$$p(y = 3, N = 10 | M_1) = \int_{0.5}^1 p(y = 3, N = 10 | \theta, M_1) p(\theta | M_1) d\theta \approx 0.029, \quad (4)$$

and

$$p(y = 3, N = 10 | M_2) = \int_0^{0.5} p(y = 3, N = 10 | \theta, M_2) p(\theta | M_2) d\theta \approx 0.097. \quad (5)$$

Given the models have equal probability under our prior, we can calculate the posterior probability of Model 1 as

$$p(M_1 | y = 3, N = 10) = \frac{p(y = 3, N = 10 | M_1)}{p(y = 3, N = 10 | M_1) + p(y = 3, N = 10 | M_2)} \approx 0.23. \quad (6)$$

This is the correct posterior probability of Model 1 given our prior and data.

To give the data more weight relative to the prior, we could use it twice, and calculate an empirical Bayes estimate using a prior of  $beta(13, 17)$ . This results in a “posterior” distribution of  $beta(16, 24)$  (Figure 1), with a MAP estimate of 0.395, and  $p(M_1 | y = 3, N = 10) = 0.10$ . The estimated posterior distribution of the parameter, and resulting MAP estimate, is similar whether or not an empirically informed prior is used. However, the posterior probability of Model 1 is very sensitive to the empirical prior, decreasing by 56%. By using the empirically informed prior, we ignored prior uncertainty, leading to an underestimate of our posterior uncertainty (Figure 1). While this did not greatly affect our estimate of  $\theta$ , it misled us to be overconfident in Model 2.

## 2.2 Practical concerns about empirically informed uniform priors for Bayesian model choice

In addition to the theoretical concerns discussed above, there are practical problems with using narrow, empirically informed, uniform priors for a method that has already been shown to be biased toward models with less parameter space. The results of Hickerson et al.’s 2014 reanalysis of the Philippine dataset strongly favored models with the narrowest, empirically informed prior on divergence time, and thus their model-averaged posterior estimates are dominated by models  $M_1$  and  $M_2$  (see Table 1 of Hickerson et al. (2014)). This is concerning,

because the narrowest  $\tau$  prior used by Hickerson et al. (2014) ( $\tau \sim U(0, 0.1)$ ) likely excludes the true divergence times for at least some of the Philippine taxa, a major problem when using uniform priors. Hickerson et al. (2014) set this prior to match the 95% highest posterior density (HPD) interval for the mean divergence time estimated under one of the priors used by Oaks et al. (2013) (see Tables 2 and 3 of Oaks et al. (2013)). Given this interval estimate is for the *mean* divergence time across all 22 taxa, it may be inappropriate to set this as the limit on the prior, because some of the taxon pairs are expected to have diverged at times older than the upper limit. Furthermore, this prior is *excluded* from the 95% HPD interval estimates of the mean divergence time under the other two priors explored by Oaks et al. (2013) (under these priors the 95% HPD is approximately 0.3–0.6; see Table 6 of Oaks et al. (2013)).

The strong preference for the questionable priors on divergence times (their  $M_1$  and  $M_2$ ) in the model-averaged results of Hickerson et al. (2014) suggest their approach is biased toward models with less parameter space, and as a consequence, is biased toward estimating model-averaged posteriors dominated by models that exclude true values of parameters of the model. We explored this possibility in two ways. First, we re-analyzed the Philippines dataset using the model-averaging approach of Hickerson et al. (2014), but set one of the prior models with a uniform prior on divergence times that is unrealistically narrow, and almost certainly excludes most, if not all, of the true divergence times of the 22 taxon pairs. If small likelihoods of large models cause the method to prefer models with less parameter space, we expect `msBayes` will preferentially sample from this incorrect model yielding a posterior that is incorrect (i.e., the model-averaged posterior will be dominated by an incorrect model that excludes the truth). Second, we generated simulated datasets for which the divergence times are drawn from an exponential distribution and applied the approach of Hickerson et al. (2014) to each of them to see how often the method excludes the truth.

### 2.2.1 Re-analyses of the Philippines dataset using empirical Bayesian model averaging

For our re-analyses of the Philippines dataset we followed the model-averaging approach of Hickerson et al. (2014), but with a reduced set of prior models to avoid their error of mixing units of time (see SI for details). We used five prior models, all of which had priors on population sizes of  $\theta_D \sim U(0.0001, 0.1)$  and  $\theta_A \sim U(0.0001, 0.05)$ . Following Hickerson et al. (2014), each of these models had the following priors on divergence time parameters:  $M_1$ ,  $\tau \sim U(0, 0.1)$ ;  $M_2$ ,  $\tau \sim U(0, 1)$ ;  $M_3$ ,  $\tau \sim U(0, 5)$ ;  $M_4$ ,  $\tau \sim U(0, 10)$ ; and  $M_5$ ,  $\tau \sim U(0, 20)$ . We simulated  $1 \times 10^6$  random samples from each of the models for a total of  $5 \times 10^6$  prior samples. For each model, we retained the 10,000 samples with the smallest Euclidean distance from the observed summary statistics, standardizing the statistics using the prior means and standard deviations of the given model. From the remaining 50,000 samples, we then retained the 10,000 samples with the smallest Euclidean distance from the observed summary statistics, this time standardizing the statistics using the prior means and standard deviations across all five models. We then repeated this analysis twice, replacing the  $M_1$  model with  $M_{1A}$  and  $M_{1B}$ , which differ only by having priors on divergence times of  $\tau \sim U(0, 0.01)$  and  $\tau \sim U(0, 0.001)$ , respectively. While we suspect the prior of  $\tau \sim U(0, 0.1)$  used by Hickerson et al. (2014) likely excludes the true divergence times of at least some of the 22 taxa, we are



nearly certain that these narrower priors are incorrect and exclude most, if not all, of the divergence times of the Philippine taxa.

Our results show that the model-averaging approach of Hickerson et al. (2014) does not reduce the method’s support of models with less parameter space. Rather, the method strongly prefers the prior model with the narrowest distribution on divergence times across all three of our analyses, even when this model is almost certainly incorrect and excludes the true divergence times of the Philippine taxa (Table 1).

However, Hickerson et al. (2014) vetted the priors used in their model-averaging approach via “graphical checks,” in which the summary statistics from 1000 random samples of each prior model are plotted along the first two orthogonal axes of a principle component analysis (see Figure 1 of Hickerson et al. (2014)). To determine if such prior predictive analyses would indicate the  $M_{1A}$  and  $M_{1B}$  models are problematic, we performed these graphical checks on our prior models. Unfortunately, these prior predictive checks provide no warning that these priors are too narrow (Figure S2). Rather, the graphs suggest these incorrect priors are “better fit” (Figure S2A–C) than the valid priors similar to those used by Oaks et al. (2013) (Figure S2D–F).

Given that the results of Hickerson et al. (2014) strongly prefer the models with the narrowest prior on divergence times, it seems quite likely that their model-averaged results are dominated by models that exclude true divergence times, making their results difficult to interpret.

### 2.2.2 Simulation-based assessment of Hickerson et al.’s 2014 model averaging over empirical priors

To better quantify the propensity of Hickerson et al.’s (2014) approach to exclude the truth, we simulated 1000 datasets in which the divergence times for the 22-population pairs are drawn randomly from an exponential distribution with a mean of 0.5 ( $\tau \sim Exp(2)$ ). All other parameters were identically distributed as the  $M_1$ – $M_5$  models (Table 1). We then repeated the analysis described above using  $1 \times 10^6$  random samples from prior models  $M_1$ ,  $M_2$ ,  $M_3$ ,  $M_4$ , and  $M_5$ , retaining 1000 posterior samples for each of the 1000 simulated datasets.

For each simulation replicate, we estimated the Bayes factor in favor of excluding the truth as the ratio of the posterior to prior odds of excluding the true value of at least one parameter. Whenever the Bayes factor preferred a model excluding the truth we counted the number of the 22 true divergence times that were excluded by the preferred model. Our results show that the model-averaging approach of Hickerson et al. (2014) favors a model that excludes the true values of parameters in 97% of the replicates (90% with GLM-regression adjustment), excluding up to 21 of the 22 true divergence times (Figure 2). We also used the mode estimate of the preferred model for each replicate to estimate the number of true parameter values excluded, which produced very similar results. Importantly, the posterior probability of excluding at least one true parameter value is very high in nearly all of the replicates (Figure 3). Using a Bayes factor of greater than 10 as a criterion for strong support, 66% of the replicates (87% with GLM-regression adjustment) strongly support the exclusion of true values (Figure 3).

The results of the preceding empirical and simulation-based analyses clearly demonstrate

the risk of using narrow, empirically guided uniform priors in a Bayesian model-averaging framework. The consequence of this approach is obtaining a model-averaged posterior estimate that is heavily weighted toward incorrect models that exclude true values of model parameters.

## 2.3 Additional thoughts on empirical priors in Bayesian model choice

The preceding sections are not a general critique of Bayesian model averaging. Rather, model averaging can provide an elegant way of incorporating model uncertainty in Bayesian inference. However, when averaging over models with narrow and broad uniform priors on a parameter with a likelihood density that is not expected to be uniformly distributed, the posterior can be dominated by models that exclude from consideration the true values of parameters due to the larger marginal likelihoods of the models that are integrated over less space with high prior weight yet low likelihood.

Given the theoretical and practical issues with empirical Bayes approaches to Bayesian model choice, it is clear why one should use caution before overly criticizing an investigator’s choice of priors *after* having seen the resulting posterior. As discussed by Oaks et al. (2013), prior to analyzing the data, there was a large amount of uncertainty regarding the divergence times of the 22 population pairs under study. Two of these pairs represent distinct species, and the taxonomy of many groups in the Philippines has repeatedly been shown to mask deeply divergent lineages (Brown et al., 2008; Linkem et al., 2010; Siler et al., 2010; Welton et al., 2010; Siler et al., 2011b,a, 2012; Brown and Stuart, 2012; Linkem and Brown, 2013; Brown et al., 2013; Siler et al., 2014). When limited to uniformly distributed priors, the alternative to priors that reflect prior uncertainty, as shown above, is to risk excluding the true values one seeks to estimate. However, continuous distributions more appropriate as priors for positive real-valued parameters have been shown to greatly reduce spurious support for clustered divergence models while allowing prior uncertainty to be accommodated (Oaks, 2014).

## 3 Assessing the power of the model-averaging approach of Hickerson et al. (2014)

While our results above clearly demonstrate the risks inherent to the empirical Bayesian model-choice approach used by Hickerson et al. (2014), one could justify such risk if the method does indeed increase the power of the method and thus decrease bias toward clustered divergence models. We assess this possibility using simulation-based analyses. Following Oaks et al. (2013), we simulated 1000 datasets with  $\tau$  for each of the 22 population pairs randomly drawn from a uniform distribution,  $U(0, \tau_{max})$ , where  $\tau_{max}$  was set to: 0.2, 0.4, 0.6, 0.8, 1.0, and 2.0, in  $4N_C$  generations. All other parameters were identically distributed as the prior models. As above, we generated a prior sample of  $5 \times 10^6$  total samples from the five prior models  $M_1$ ,  $M_2$ ,  $M_3$ ,  $M_4$ , and  $M_5$  (Table 1). For each simulated dataset, we approximated the posterior by retaining 1000 samples from the prior with the smallest Euclidean distance from the true summary statistics, as described above. In total, we analyzed 6000 replicate datasets, retaining 1000 model-averaged posterior samples for each of them.

We find that the approach of Hickerson et al. (2014) struggles to estimate the variance of divergence times ( $\Omega$ ) across most of the  $\tau_{max}$  we simulated, whether evaluating the unadjusted (Figure S3A–F) or GLM-adjusted (Figure S3G–L) posterior estimates. The method only estimates  $\Omega$  relatively well when the simulated distribution of divergence times is identical to one of the prior models (Figures S3E&K). This is consistent with the findings of Oaks et al. (2013) and Hickerson et al. (2014) that **msBayes** is highly sensitive to the prior distribution deviating from the true, underlying distribution of the data.

Furthermore, our results demonstrate that the approach of Hickerson et al. (2014) consistently infers highly clustered divergences across all the  $\tau_{max}$  we simulated (Figure 4A–D & S4A–F). The method is most likely to infer the extreme case of a single divergence event shared across 22 taxa when populations diverged randomly over the past  $4N_C$  generations. When divergences are random over the past  $8N_C$  generations, the most likely inference is only two divergence events, and a single divergence is still estimated in more than 10% of the replicates. It is interesting to note that as  $\tau_{max}$  increases, but before the estimates are finally pulled away from  $\Psi = 1$ , the distribution of  $\Psi$  estimates closely mirror the U-shaped prior on divergence models used by **msBayes** (see Figure 4E and Oaks et al.’s (2013) Figure 5B). This suggests this U-shaped prior coupled with the small marginal likelihoods of models with many divergence parameters, is a major cause of the method’s bias toward clustered divergence models; this is consistent with Oaks et al. (2013) and confirmed by Oaks (2014).

Looking at our simulation results in terms of the posterior probability of the dispersion index of divergence times supporting the extreme case of one divergence event (i.e.,  $p(\Omega < 0.01 | B_\epsilon(\mathbf{S}^*))$ ), we find the method strongly supports one divergence in greater than 27% of the replicates across all the  $\tau_{max}$  we simulated (Figure 4E–H & S4G–L); following Hickerson et al. (2014), we use a Bayes factor of greater than 10 as the criterion for incorrect inference of a single divergence event. Furthermore, there is strong support for a single divergence event in more than 90% of the replicates when divergences are random over the past  $2.4N_C$  generations, and more than 60% when over the past  $3.2N_C$  generations (Figure 4E–H & S4G–L).

Our results show that the empirical Bayesian model-averaging approach leads to spuriously strong support for the extreme case of a single divergence event when populations diverged randomly over the last  $8N_C$  generations. To put this on the scale roughly consistent with a vertebrate mitochondrial locus, assuming a mutation rate of  $2 \times 10^{-8}$  per site per generation, this translates to 5 million generations. Assuming a mutation rate consistent with nuclear loci of  $1 \times 10^{-9}$ , this is 100 million generations. Given the empirically Bayesian model-averaging approach of Hickerson et al. (2014) lacks power to detect random variation in divergence times over such timescales, it is difficult to justify the risk of the approach to exclude regions of parameter space containing the truth.

Also, the results of our power analyses further demonstrate the propensity of Hickerson et al.’s (2014) approach to exclude true parameter values. Across all but one of the  $\tau_{max}$  we simulated, in a large proportion of replicates the method favors a model that excludes the truth, and across many of the  $\tau_{max}$  the preferred model will exclude a large proportion of the true divergence-time values (Figure 5A–D & S5A–F). Importantly, the posterior probability of excluding at least one true divergence value is also quite high across many of the  $\tau_{max}$  (Figure 5E–H & S5G–L). Only when the data are identically distributed as one of the prior models does the method avoid excluding the truth more than 5% of the time (Figure 5C).

Again, this demonstrates the method’s sensitivity to priors.

## 4 The importance of power analysis to guide applications of msBayes

Hickerson et al. (2014) presented a power analysis of **msBayes** under a narrow uniform divergence-time prior of 0–1 coalescent units ago. They found that under these prior conditions **msBayes** can, assuming a per-site rate of  $1.92 \times 10^{-8}$  mutations per generation, detect multiple divergence events among 18 taxa when the true divergences were random over hundreds of thousands of generations or more. Oaks et al. (2013) performed similar power analyses under three uniform divergence-time priors as narrow as 0–5 coalescent units, and found the method was able to detect multiple events among 22 when divergences were random over millions of generations. As recommended by Oaks et al. (2013), it is important that investigators perform power analyses to determine the method’s power for their dataset, and decide if **msBayes** has sufficient temporal resolution to address their hypotheses; in the case of the Philippines dataset, it did not. When doing so, it is important to consider what prior conditions are relevant to their empirical system. A divergence-time prior of 0–5 coalescent units is quite narrow considering that this expresses the prior belief that all 22 taxon pairs diverged within this window. Certainly, it is rare for there to be enough *a priori* information to be certain that all taxa diverged within the last  $4N_C$  generations (i.e., 0–1 coalescent units). Also, it seems unlikely that when such prior information is available that being able to detect multiple divergences on the scale of hundreds of thousands of generations will add much insight about the evolutionary history of the taxa; assuming a rate of mutation consistent with nuclear loci ( $1 \times 10^{-9}$ ), this translates to a temporal resolution of 3 million generations or more.

As noted by Oaks et al. (2013), inferring more than one divergence time shared across all taxa does not confirm the method is working well when analyzing data generated under random temporal variation in divergences (e.g., an inference of two divergence events could be biogeographically interesting yet spurious). Thus, it is important that investigators not limit their assessment of the method’s power to only differentiating inferences of one event or more (i.e.  $\Psi = 1$  versus  $\Psi > 1$ ). Rather, looking at the distribution of estimates, as in Oaks et al. (2013), provides much more information about the behavior of the method.

## 5 The causes of support for models of co-divergence

To determine how best to improve the behavior of **msBayes**, it is important to determine the mechanism by which broad uniform priors cause spurious support for clustered models of divergence. It is well established that vague priors can be problematic in Bayesian model selection. Models that integrate over more parameter space characterized by low probability of producing the data and relatively high prior density will have smaller marginal likelihoods (Jeffreys, 1939; Lindley, 1957). Given the uniformly distributed priors on divergence times (and other nuisance parameters) employed in **msBayes**, models with more divergence parameters will be forced to integrate over *much* greater parameter space, all with equal prior

density, and much of it with low likelihood. In light of this fundamental statistical issue, it is not surprising that the method tends to support simple models.

However, Hickerson et al. (2014) conclude that the bias is due to insufficient sampling. They argue the widest of the three priors used by Oaks et al. (2013) would infrequently produce samples with many independent population divergence times as recent as the estimated gene divergence times presented in Oaks et al. (2013). However, this argument assumes the gene divergence times presented in Oaks et al. (2013) were estimated without error. These estimates were intended to only provide a rough comparison of the gene divergence times across the 22 taxa. These analyses assumed an arbitrary strict mutation rate of  $2 \times 10^{-8}$  for all taxa, and are, of course, subject to estimation error. Furthermore, the branch-length units of the gene trees are in millions of years, whereas the divergence time prior of **msBayes** is in generations, thus Hickerson et al. (2014) make the implicit assumption that all 22 Philippine taxa have a generation time of one year. The argument of Hickerson et al. (2014) that divergence time estimates of Oaks et al. (2013) “should set an upper bound on their prior for  $\tau$ ” seems questionable, especially given our findings presented above regarding the behavior of **msBayes** when empirically informed uniform priors are employed.

Even if we assume the gene divergence times are estimated without error, Hickerson et al.’s 2014 argument only applies to one of the three different priors used by Oaks et al. (2013). The narrowest prior on divergence times used by Oaks et al. (2013) ( $U(0, 5)$ ) closely mirrors the range of estimates of gene-divergence times if we assume the same mutation rate and one generation per year (0–5 million years ago). Applying Hickerson et al.’s (2014) sampling-probability argument demonstrates this prior is densely populated with samples with large numbers of divergence parameters with values younger than the estimated gene divergence estimates. Thus, if insufficient prior sampling is to blame for the bias, it should be much reduced under the narrow prior on  $\tau$ . However, the magnitude of the bias is very similar across all three priors explored by Oaks et al. (2013). Hickerson et al. (2014) point out a case where the narrow prior performs slightly better (panel L of Figures S32, S37, and S38 of Oaks et al. (2013)). However, it is important to note that these results suffered from a bug in **msBayes**, and there are many cases after Oaks et al. (2013) corrected the bug where the narrow prior performs slightly worse (see panels D–J of Figures 3 and S12).

To disentangle whether model likelihoods or insufficient prior sampling is to blame for the method’s spurious support for simple models, we must look at the different predictions made by these two phenomena. One example, as discussed by Oaks et al. (2013), is that insufficient prior sampling should create large variance among posterior estimates, and thus it should cause analyses to be highly sensitive to the number of samples drawn from the prior. Furthermore, if sampling error is biased toward models with less parameter space, as suggested by Hickerson et al. (2014), we expect to see support for these models decrease as sampling increases. Oaks et al. (2013) did not see such sensitivity when they compared prior sample sizes of  $2 \times 10^6$ ,  $5 \times 10^6$ , and  $10^7$ .

To explore this prediction further, we repeat the analysis of the Philippines dataset under the intermediate prior used by Oaks et al. (2013) ( $\tau \sim U(0, 10)$ ,  $\theta_D \sim (0.0005, 0.04)$ ,  $\theta_A \sim (0.0005, 0.02)$ ), using a very large prior sample size of  $10^8$ . When we look at the trace of the estimates of the dispersion index of divergence times ( $\Omega$ ) as the prior samples accumulate (Figure 6) we do not see the trend predicted by Hickerson et al. (2014) in either the unadjusted or GLM-regression-adjusted estimates. While sampling error is always a

reality, it does not appear to be playing a large role in the biases revealed by the results of Oaks et al. (2013) or presented above.

As discussed by Oaks et al. (2013), a straightforward prediction if marginal likelihoods are causing the preference for simple models is that the bias should disappear as the model generating the data converges to the prior. Oaks et al. (2013) tested this prediction by performing 100,000 simulations to assess the model-choice behavior of **msBayes** when the prior model is correct. The results confirm the prediction as **msBayes** actually tends to underestimate the probability of the one-divergence model (see Figure 4 of Oaks et al. (2013)). We confirmed this same behavior for the model-averaging approach used by Hickerson et al. (2014) (see SI text and Figure S6). These results are not clearly predicted if insufficient sampling was causing the bias. Even when the prior is correct, due to the discrete uniform prior on  $\Psi$  implemented in **msBayes**, models with larger numbers of divergence events (and thus greater parameter space) will still be less densely sampled than those with fewer divergence events (Oaks et al., 2013). Thus, the results of the simulations of Oaks et al. (2013) are more consistent with the fundamental sensitivity of marginal likelihoods to priors.

This is further demonstrated by the results presented herein that show the model-averaging approach of Hickerson et al. (2014) prefers models with narrower  $\tau$  priors (Table 1 and Figs. 2, 3 and 5) and models with fewer  $\tau$  parameters (Figure 4). In all of these analyses, each of the prior models have the same number of samples. Thus, while sampling error will always be present in any numerical Bayesian approximation method, insufficient sampling is an unlikely explanation for the support of prior models with less parameter space. While analyses that sample each model proportional to their parameter space could be explored, it is clear that the marginal likelihoods under broad uniform priors on divergence times will be greater for models with fewer divergence-time dimensions.

## 6 General thoughts on the model of **msBayes**

Our findings are not surprising in light of the difficult inference problem with which **msBayes** is faced. To get a sense of the difficulty of this problem, we tally up all the free parameters of the **msBayes** model as applied to the dataset of Oaks et al. (2013). Under the simplest model in **msBayes** (i.e., assuming no migration and no intra-locus recombination), the number of parameters for each taxon pair include: The population sizes of the ancestral and descendant populations ( $\theta_A$ ,  $\theta_{D1}$ ,  $\theta_{D2}$ ), the magnitude of population contraction in each of the descendant populations ( $\zeta_{D1}$  and  $\zeta_{D2}$ ), the timing of these contractions ( $\tau_B$ ), and the  $N - 1$  node heights (coalescent times) of the gene tree that gave rise to the  $N$  gene copies sampled from both populations. Lastly, there are between one and 22 divergence time parameters  $\tau$  in the vector  $\boldsymbol{\tau}$ . Overall, when applying the simplest model in **msBayes** to the dataset of Oaks et al. (2013) with 22 taxon pairs, there are 581–602 free parameters (depending on the number of divergence-time parameters in  $\boldsymbol{\tau}$ ). Furthermore, under this rich model, the method is estimating the probability of 1002 divergence models (i.e., the number of integer partitions of  $Y = 22$ ; Oaks et al., 2013).

This is a very difficult inference problem, especially considering that all the information in the sequence alignment of each taxon pair is distilled into four summary statistics:  $\pi$  (Tajima, 1983),  $\theta_W$  (Watterson, 1975),  $\pi_{net}$  (Takahata and Nei, 1985), and  $SD(\pi - \theta_W)$

(Tajima, 1989). That gives us a total of 88 summary statistics (four from each of the 22 taxon pairs), which contain minimal information about many of the  $\approx 600$  parameters in the model. More summary statistics can be used in `msBayes`, but most are highly correlated with the four default statistics, and thus contribute little additional information about the parameters from the sequence data. The large number of parameters and divergence models relative to the amount of information in the data is undoubtedly a key reason the method lacks robustness to prior conditions. Robustness is an important characteristic of a method to gauge its applicability to real-world data, because we know the model and priors will be wrong to some degree.

The model-averaging approach of Hickerson et al. (2014) adds an additional dimension of model choice to the model. They expand the model to sample over eight prior models. This adds an additional free parameter to the model and, more importantly, forces the model to sample over 8016 unique models. While this approach is trivial to implement, it is a non-trivial extension of the model. In theory, the model-averaging approach of Hickerson et al. (2014) is very appealing. It leverages a great strength of Bayesian statistical procedures, namely the ability to obtain marginalized estimates that incorporate uncertainty in nuisance parameters. However, when averaging over both narrow-empirical and diffuse uniform priors for a parameter that is expected to have a very non-uniform likelihood density, and in the context of a model-choice method that is highly sensitive to priors, it is not too surprising that the approach struggles.

The recommendations of Oaks et al. (2013) for mitigating the lack of robustness of `msBayes` are similar to those of Hickerson et al. (2014), but avoid the need for imposing an additional dimension of model choice. Oaks et al. (2013) suggest that uniform priors may not be appropriate for many parameters of the `msBayes` model, and recommend the use of probability distributions from the exponential family. If we look at the prior distribution on divergence-time parameters imposed by the model-averaging approach of Hickerson et al. (2014) we see it is a mixture of overlapping uniforms with lower limits of zero (Figure 7). This looks very much like an exponential distribution, except that in any state of the model, all the divergence-time parameters are restricted to the hard bounds of one of the uniform distributions. Thus, it seems more appropriate to simply place a gamma prior (the exponential being a special case of the gamma) on divergence times. This would capture the prior uncertainty that Hickerson et al. (2014) are suggesting for divergence times (Figure 7) while avoiding costly model-averaging and the constraint that all divergence times must fall within the hard bounds of the current model state. It also would allow an investigator to place the majority of the prior density in regions of parameter space they believe, *a priori*, are most plausible, but still capture uncertainty in the tails of distributions with low density. Most importantly, we have found the use of more flexible distributions in place of uniform priors improves the power of the method to detect temporal variation in divergences and reduces the bias toward models of clustered divergences (Oaks, 2014).

## 7 Conclusions

We demonstrate how the approximate Bayesian model-choice method implemented in `msBayes` can be strongly biased away from models with greater parameter space. As sug-

gested by Oaks et al. (2013), this is likely caused by the use of uniform priors on most of the model’s parameters. Uniform distributions necessitate the use of broad priors that place high prior density in unlikely regions of parameter space, less the risk of excluding the truth *a priori*. These broad priors reduce the marginal likelihoods of models with more divergence-time parameters or broader prior distributions on those parameters. We show that the empirical Bayesian model-averaging approach of Hickerson et al. (2014) does not mitigate this bias, but rather causes it to manifest by sampling predominantly from models that may exclude the true values of the parameters.

Whether or not one chooses to use empirically informed priors, our results suggest that it is difficult to choose an uniformly distributed prior on divergence times that is broad enough to confidently contain the true values of parameters while being narrow enough to avoid bias toward models with less parameter space. While we do not assume “that all previous **msBayes** results are invalid,” as suggested by Hickerson et al. (2014), we do conservatively recommend that the common inference of temporally clustered divergences (Barber and Klicka, 2010; Bell et al., 2012; Carnaval et al., 2009; Chan et al., 2011; Daza et al., 2010; Hickerson et al., 2006; Huang et al., 2011; Lawson, 2010; Leaché et al., 2007; Plouviez et al., 2009; Stone et al., 2012; Voje et al., 2009), when not accompanied with the necessary analyses to assess the robustness and temporal resolution of such results, should be treated with caution, because the method has been shown to spuriously infer clustered divergences over a range of prior conditions. Fortunately, alternative prior probability distributions allow prior uncertainty to be accommodated while avoiding excessive prior density in regions of low likelihood, which greatly improves the power of the method Oaks (2014).

The work presented herein follows the principles of Open Notebook Science. All aspects of the work were recorded in real-time via version-control software and are publicly available at <https://github.com/joaks1/msbayes-experiments>. All information necessary to reproduce our results is provided there.

## 8 Acknowledgments

We thank Melissa Callahan, Jake Esselstyn, Cameron Siler, Mark Holder, Rafe Brown, Emily McTavish, Daniel Money, Jordan Koch, and Adam Leaché for insightful comments that greatly improved this work. We thank Michael Hickerson and co-authors for generously providing their data. J. Oaks and C. Linkem thank the National Science Foundation for supporting this work (DEB 1011423, DBI 1308885 and BIO-1202754). J. Oaks was also supported by the University of Kansas (KU) Office of Graduate Studies, Society of Systematic Biologists, Sigma Xi Scientific Research Society, KU Department of Ecology and Evolutionary Biology, and the KU Biodiversity Institute. We also thank Mark Holder, the KU Information and Telecommunication Technology Center, KU Computing Center, and the iPlant Collaborative for the computational support necessary to conduct the analyses presented herein.



## References

- Barber, B. R. and J. Klicka, 2010. Two pulses of diversification across the Isthmus of Tehuantepec in a montane Mexican bird fauna. *Proceedings Of The Royal Society B-Biological Sciences* 277:2675–2681.
- Bell, R. C., J. B. MacKenzie, M. J. Hickerson, K. L. Chavarria, M. Cunningham, S. Williams, and C. Moritz, 2012. Comparative multi-locus phylogeography confirms multiple vicariance events in co-distributed rainforest frogs. *Proceedings Of The Royal Society B-Biological Sciences* 279:991–999.
- Brown, R. M., A. C. Diesmos, and A. C. Alcala, 2008. Philippine amphibian biodiversity is increasing in leaps and bounds. Pp. 82–83, *in* S. N. Stuart, M. Hoffmann, J. S. Chanson, N. A. Cox, R. Berridge, P. Ramani, and B. E. Young, eds. *Threatened Amphibians of the World*. Lynx Ediciones, Barcelona, Spain; IUCN—The World Conservation Union, Gland, Switzerland; and Conservation International, Arlington, Virginia, USA.
- Brown, R. M., C. D. Siler, C. H. Oliveros, J. A. Esselstyn, A. C. Diesmos, P. A. Hosner, C. W. Linkem, A. J. Barley, J. R. Oaks, M. B. Sanguila, L. J. Welton, R. G. Moyle, A. T. Peterson, and A. C. Alcala, 2013. Evolutionary processes of diversification in a model island archipelago. *Annual Review of Ecology, Evolution, and Systematics* 44:411–435.
- Brown, R. M. and B. L. Stuart, 2012. Patterns of biodiversity discovery through time: an historical analysis of amphibian species discoveries in the Southeast Asian mainland and island archipelagos. Pp. 348–389, *in* D. J. Gower, K. G. Johnson, J. E. Richardson, B. R. Rosen, L. Rüber, and S. T. Williams, eds. *Biotic Evolution and Environmental Change in Southeast Asia*. Cambridge University Press.
- Carlin, B. P. and A. E. Gelfand, 1990. Approaches for empirical Bayes confidence intervals. *Journal of the American Statistical Association* 85:105–114.
- Carnaval, A. C., M. J. Hickerson, C. F. B. Haddad, M. T. Rodrigues, and C. Moritz, 2009. Stability Predicts Genetic Diversity in the Brazilian Atlantic Forest Hotspot. *Science* 323:785–789.
- Chan, L. M., J. L. Brown, and A. D. Yoder, 2011. Integrating statistical genetic and geospatial methods brings new power to phylogeography. *Molecular Phylogenetics and Evolution* 59:523–537.
- Daza, J. M., T. A. Castoe, and C. L. Parkinson, 2010. Using regional comparative phylogeographic data from snake lineages to infer historical processes in Middle America. *Ecography* 33:343–354.
- Efron, B., 2008. Microarrays, Empirical Bayes and the Two-Groups Model. *Statistical Science* 23:1–22.
- , 2013. Empirical bayes modeling, computation, and accuracy. Manuscript AMS 2010 subject classifications: Primary 62C10; secondary 62-07, 62P10.

- Hickerson, M. J., E. A. Stahl, and H. A. Lessios, 2006. Test for simultaneous divergence using approximate Bayesian computation. *Evolution* 60:2435–2453.
- Hickerson, M. J., G. N. Stone, K. Lohse, T. C. Demos, X. Xie, C. Landerer, and N. Takebayashi, 2014. Recommendations for using msbayes to incorporate uncertainty in selecting an ABC model prior: A response to Oaks et al. *Evolution* 68:248–294.
- Huang, W., N. Takebayashi, Y. Qi, and M. J. Hickerson, 2011. MTML-msBayes: Approximate Bayesian comparative phylogeographic inference from multiple taxa and multiple loci with rate heterogeneity. *BMC Bioinformatics* 12:1.
- Hwang, J. T. G., J. Qiu, and Z. Zhao, 2009. Empirical Bayes confidence intervals shrinking both means and variances. *Journal of the Royal Statistical Society Series B-Statistical Methodology* 71:265–285.
- Jeffreys, H., 1939. *Theory of Probability*. 1st ed. Clarendon Press, Oxford, U.K.
- Laird, N. M. and T. A. Louis, 1987. Empirical Bayes confidence intervals based on bootstrap samples. *Journal of the American Statistical Association* 82:739–750.
- , 1989. Empirical Bayes confidence intervals for a series of related experiments. *Biometrics* 45:481–495.
- Lawson, L. P., 2010. The discordance of diversification: evolution in the tropical-montane frogs of the Eastern Arc Mountains of Tanzania. *Molecular Ecology* 19:4046–4060.
- Leaché, A. D., S. C. Crews, and M. J. Hickerson, 2007. Two waves of diversification in mammals and reptiles of Baja California revealed by hierarchical Bayesian analysis. *Biology Letters* 3:646–650.
- Lindley, D. V., 1957. A statistical paradox. *Biometrika* 44:187–192.
- Linkem, C. W. and R. M. Brown, 2013. Systematic revision of the *Parvosцинus decipiens* (Boulenger, 1894) complex of Philippine forest skinks (Squamata: Scincidae: Lygosominae) with descriptions of seven new species. *Zootaxa* 3700:501–533.
- Linkem, C. W., K. M. Hesed, A. C. Diesmos, and R. M. Brown, 2010. Species boundaries and cryptic lineage diversity in a Philippine forest skink complex (Reptilia; Squamata; Scincidae: Lygosominae). *Molecular Phylogenetics and Evolution* 56:572–585.
- Morris, C. N., 1983. Parametric empirical bayes inference: Theory and applications. *Journal of the American Statistical Association* 78:47–55.
- Oaks, J. R., 2014. An improved approximate-bayesian model-choice method for estimating shared evolutionary history. [arXiv:1402.6303 \[q-bio.PE\]](https://arxiv.org/abs/1402.6303) .
- Oaks, J. R., J. Sukumaran, J. A. Esselstyn, C. W. Linkem, C. D. Siler, M. T. Holder, and R. M. Brown, 2013. Evidence for climate-driven diversification? a caution for interpreting ABC inferences of simultaneous historical events. *Evolution* 67:991–1010.

- Plouviez, S., T. M. Shank, B. Faure, C. Daguin-Thiebaut, F. Viard, F. H. Lallier, and D. Jollivet, 2009. Comparative phylogeography among hydrothermal vent species along the East Pacific Rise reveals vicariant processes and population expansion in the South. *Molecular Ecology* 18:3903–3917.
- Siler, C. D., A. C. Diesmos, A. C. Alcala, and R. M. Brown, 2011a. Phylogeny of Philippine slender skinks (Scincidae: *Brachymeles*) reveals underestimated species diversity, complex biogeographical relationships, and cryptic patterns of lineage diversification. *Molecular Phylogenetics and Evolution* 59:53–65.
- Siler, C. D., A. M. Fuiten, R. M. Jones, A. C. Alcala, and R. M. Brown, 2011b. Phylogeny-based species delimitation in Philippine slender skinks (Reptilia: Squamata: Scincidae) II: Taxonomic revision of *Brachymeles samarensis* and description of five new species. *Herpetological Monographs* 25:76–112.
- Siler, C. D., J. R. Oaks, K. Cobb, O. Hidetoshi, and R. M. Brown, 2014. Critically endangered island endemic or peripheral population of a widespread species? conservation genetics of Kikuchi’s gecko and the global challenge of protecting peripheral oceanic island endemic vertebrates. *Diversity and Distributions* .
- Siler, C. D., J. R. Oaks, J. A. Esselstyn, A. C. Diesmos, and R. M. Brown, 2010. Phylogeny and biogeography of Philippine bent-toed geckos (Gekkonidae: *Cyrtodactylus*) contradict a prevailing model of Pleistocene diversification. *Molecular Phylogenetics and Evolution* 55:699–710.
- Siler, C. D., J. R. Oaks, L. J. Welton, C. W. Linkem, J. Swab, A. C. Diesmos, and R. M. Brown, 2012. Did geckos ride the Palawan raft to the Philippines? *Journal of Biogeography* 39:1217–1234.
- Stone, G. N., K. Lohse, J. A. Nicholls, P. Fuentes-Utrilla, F. Sinclair, K. Schönrogge, G. Csóka, G. Melika, J.-L. Nieves-Aldrey, J. Pujade-Villar, M. Tavakoli, R. R. Askew, and M. J. Hickerson, 2012. Reconstructing community assembly in time and space reveals enemy escape in a Western Palearctic insect community. *Current Biology* 22:532–537.
- Tajima, F., 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105:437–460.
- , 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123:585–595.
- Takahata, N. and M. Nei, 1985. Gene genealogy and variance of interpopulational nucleotide differences. *Genetics* 110:325–344.
- Voje, K. L., C. Hemp, Ø. Flagstad, G.-P. Saetre, and N. C. Stenseth, 2009. Climatic change as an engine for speciation in flightless Orthoptera species inhabiting African mountains. *Molecular Ecology* 18:93–108.
- Watterson, G. A., 1975. On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology* 7:256–276.

Welton, L. J., C. D. Siler, C. W. Linkem, A. C. Diesmos, and R. M. Brown, 2010. Philippine bent-toed geckos of the *Cyrtodactylus agusanensis* complex: Multilocus phylogeny, morphological diversity, and descriptions of three new species. *Herpetological Monographs* 24:55–85.

## Figure Captions

- Figure 1. A plot of three beta probability density functions that represent a prior (black;  $\text{beta}(10, 10)$ ), posterior (blue;  $\text{beta}(13, 17)$ ), and empirical Bayes density (red;  $\text{beta}(16, 24)$ ) for a dataset of 10 coin flips, three of which are successes.
- Figure 2. Histograms of the number of true divergence times excluded from the model preferred by the empirically informed model-averaging approach of Hickerson et al. (2014) when applied to simulated datasets in which divergence times of 22 pairs of populations are drawn from an exponential distribution,  $\tau \sim \text{Exp}(2)$ . The plots represent (A) unadjusted and (B) GLM-adjusted estimates from 1000 simulation replicates analyzed using  $5 \times 10^6$  samples from the prior. The proportion of simulation replicates in which at least one true parameter value is excluded from the preferred model ( $p(\tau \notin \hat{M})$ ) is also given.
- Figure 3. Histograms of the support (estimated posterior probabilities) for excluding at least one true divergence time when the empirically informed model-averaging approach of Hickerson et al. (2014) is applied to simulated datasets in which divergence times of 22 pairs of populations are drawn from an exponential distribution,  $\tau \sim \text{Exp}(2)$ . The plots represent (A) unadjusted and (B) GLM-adjusted estimates from 1000 simulation replicates analyzed using  $5 \times 10^6$  samples from the prior. The proportion of simulation replicates in which there is strong support for at least one true parameter value being excluded from the model ( $p(BF_{\tau \notin M, \tau \in M} > 10)$ ) is also given.
- Figure 4. The tendency of the empirically informed model-averaging approach of Hickerson et al. (2014) to (A–D) infer clustered divergences and (E–H) support the extreme model of one divergence when applied to simulated datasets in which the divergence times of 22 pairs of populations are randomly drawn from the uniform distributions  $\tau \sim U(0, \tau_{max})$  indicated at the top of each column of plots (divergence-time distributions are given in units of millions of generations ago (MGA) assuming a per-site rate of  $1 \times 10^{-8}$  mutations per generation). Four of the six  $\tau_{max}$  we simulated are provided; please see Figure S4 for a summary of all of the results.

- Figure 5. Histograms of the (A–D) number of true divergence-time parameters excluded from the preferred model and the (E–H) posterior probability of excluding at least one divergence-time parameter when the empirically informed model-averaging approach of Hickerson et al. (2014) is applied to simulated datasets in which divergence times of 22 pairs of populations are randomly drawn from the uniform distributions  $\tau \sim U(0, \tau_{max})$  indicated at the top of each column of plots (divergence-time distributions are given in units of millions of generations ago (MGA) assuming a per-site rate of  $1 \times 10^{-8}$  mutations per generation). Four of the six  $\tau_{max}$  we simulated are provided; please see Figure S5 for a summary of all of the results.
- Figure 6. Traces of the estimated lower and upper limits of the 95% highest posterior density (HPD) interval of  $\Omega$  (the dispersion index of divergence times) as 100 million prior samples are accumulated. Each pair of points is based on 1000 posterior samples retained from the prior. Both (A) unadjusted and (B) GLM-regression-adjusted estimates are shown. The data analyzed were the 22 pairs of Philippine taxa from Oaks et al. (2013). Prior settings were  $\tau \sim U(0, 10)$ ,  $\theta_D \sim U(0.0005, 0.04)$ , and  $\theta_A \sim U(0.0005, 0.02)$ .
- Figure 7. The prior distribution on divergence times imposed by the model-averaging prior comprised of five models with different uniform priors on  $\tau$ :  $M_1$  ( $\tau \sim U(0, 0.1)$ ),  $M_2$  ( $\tau \sim U(0, 1)$ ),  $M_3$  ( $\tau \sim U(0, 5)$ ),  $M_4$  ( $\tau \sim U(0, 10)$ ),  $M_5$  ( $\tau \sim U(0, 20)$ ).
- Figure S1. The joint posterior of the mean ( $E(\tau)$ ) and dispersion index ( $\Omega = Var(\tau)/E(\tau)$ ) of divergence times for 22 vertebrate taxon pairs as estimated by Hickerson et al. (2014) (see Figure 2B of Hickerson et al. (2014)). The posterior samples are color-coded to indicate the erroneous mixture of timescales in the analysis of Hickerson et al. (2014); grey =  $0.05/\mu$  generations and black =  $0.02/\mu$  generations.
- Figure S2. The prior predictive graphical checks recommended by Hickerson et al. (2014) for six prior models: (A)  $M_1$  ( $\tau \sim U(0, 0.1)$ ), (B)  $M_{1A}$  ( $\tau \sim U(0, 0.01)$ ), (C)  $M_{1B}$  ( $\tau \sim U(0, 0.001)$ ), (D)  $M_3$  ( $\tau \sim U(0, 5)$ ), (E)  $M_4$  ( $\tau \sim U(0, 10)$ ), and (F)  $M_5$  ( $\tau \sim U(0, 20)$ ). The three models that likely exclude true values of some divergence times of the 22 pairs of Philippine vertebrate taxa (A–C) appear to have a better “fit” than the priors that likely cover the true divergence times (D–F). The plots project the summary statistics from 1000 random samples from each model onto the first two orthogonal axes of a principle component analysis, with the blue dot representing the observed summary statistics from the 22 population pairs of Philippine vertebrates.

Figure S3. The accuracy of (A–F) unadjusted and (G–L) GLM-adjusted estimates of dispersion index of divergence times ( $\Omega$ ) when the empirically informed model-averaging approach of Hickerson et al. (2014) is applied to simulated datasets in which divergence times of 22 pairs of populations are randomly drawn from the uniform distributions  $\tau \sim U(0, \tau_{max})$  indicated at the top of each column of plots (divergence-time distributions are given in units of millions of generations ago (MGA) assuming a per-site rate of  $1 \times 10^{-8}$  mutations per generation).

Figure S4. The tendency of the empirically informed model-averaging approach of Hickerson et al. (2014) to (A–F) infer clustered divergences and (G–L) support the extreme model of one divergence when applied to simulated datasets in which the divergence times of 22 pairs of populations are randomly drawn from the uniform distributions  $\tau \sim U(0, \tau_{max})$  indicated at the top of each column of plots (divergence-time distributions are given in units of millions of generations ago (MGA) assuming a per-site rate of  $1 \times 10^{-8}$  mutations per generation).

Figure S5. Histograms of the (A–F) number of true divergence-time parameters excluded from the preferred model and the (G–L) posterior probability of excluding at least one divergence-time parameter when the empirically informed model-averaging approach of Hickerson et al. (2014) is applied to simulated datasets in which divergence times of 22 pairs of populations are randomly drawn from the uniform distributions  $\tau \sim U(0, \tau_{max})$  indicated at the top of each column of plots (divergence-time distributions are given in units of millions of generations ago (MGA) assuming a per-site rate of  $1 \times 10^{-8}$  mutations per generation).

Figure S6. An assessment of the approximate Bayesian model-averaging approach of Hickerson et al. (2014) under the ideal conditions when the prior model is correct (i.e., the datasets are simulated from parameters drawn from the same prior distributions used in the analysis). The plots show the relationship between the estimated posterior and true probability of (A & C)  $\Psi = 1$  and (B & D)  $\Omega < 0.01$ , based on 50,000 simulations. The results summarize the (A & B) unadjusted and (C & D) GLM-adjusted posterior estimate from each simulation replicate. The prior settings for all replicates included five prior models with  $\theta_D \sim U(0.0001, 0.1)$  and  $\theta_A \sim U(0.0001, 0.05)$  for all five models, and  $M_1 : \tau \sim U(0, 0.1)$ ,  $M_2 : \tau \sim U(0, 1)$ ,  $M_3 : \tau \sim U(0, 5)$ ,  $M_4 : \tau \sim U(0, 10)$ , and  $M_5 : \tau \sim U(0, 20)$ . The number of samples from the prior was  $2.5 \times 10^6$ . The simulated data structure was 8 population pairs, with a single 1000 bp locus sampled from 10 individuals from each population. The 50,000 estimates of the posterior probability of one divergence event were assigned to 20 bins of width 0.05. The estimated posterior probability of each bin is plotted against the proportion of replicates in that bin with a true value consistent with one divergence event (i.e.,  $\Psi = 1$  or  $\Omega < 0.01$ ).

Figure S7. The summary statistics  $\pi$  (Tajima, 1983) and  $\pi_{net}$  (Takahata and Nei, 1985) as a function of divergence time between populations. Each plot represents 1100 pairs of parameter draws and summary statistics calculated from the simulated data. Prior settings for the simulations were  $\tau \sim U(0, 20)$ ,  $\theta_D \sim U(0.0005, 0.04)$ , and  $\theta_A \sim U(0.0005, 0.02)$ .



Table 1: Results of the model-averaging approach of Hickerson et al. (2014) applied to the Philippines dataset of Oaks et al. (2013) using three sets of prior models. All models used priors on population size of  $\theta_D \sim U(0.0001, 0.1)$  and  $\theta_A \sim U(0.0001, 0.05)$ , and differ only in their prior on divergence time ( $\tau$ ) parameters. Each set of five models differ only in the divergence time prior used for the model with the narrowest prior:  $M_1$  ( $\tau \sim U(0, 0.1)$ ),  $M_{1A}$  ( $\tau \sim U(0, 0.01)$ ), or  $M_{1B}$  ( $\tau \sim U(0, 0.001)$ ). The approximate posterior probability of each model ( $p(M_i | B_\epsilon(\mathbf{S}^*))$ ) is given for each of the three analyses. The posterior estimates are based on 10,000 samples retained from  $1 \times 10^6$  prior samples from each model.

Model	$\tau$ prior	$p(M_i   B_\epsilon(\mathbf{S}^*))$		
		$M_* = M_1$	$M_* = M_{1A}$	$M_* = M_{1B}$
$M_*$	–	0.899	0.821	0.673
$M_2$	$U(0, 1)$	0.079	0.136	0.251
$M_3$	$U(0, 5)$	0.013	0.026	0.044
$M_4$	$U(0, 10)$	0.006	0.012	0.022
$M_5$	$U(0, 20)$	0.003	0.005	0.010

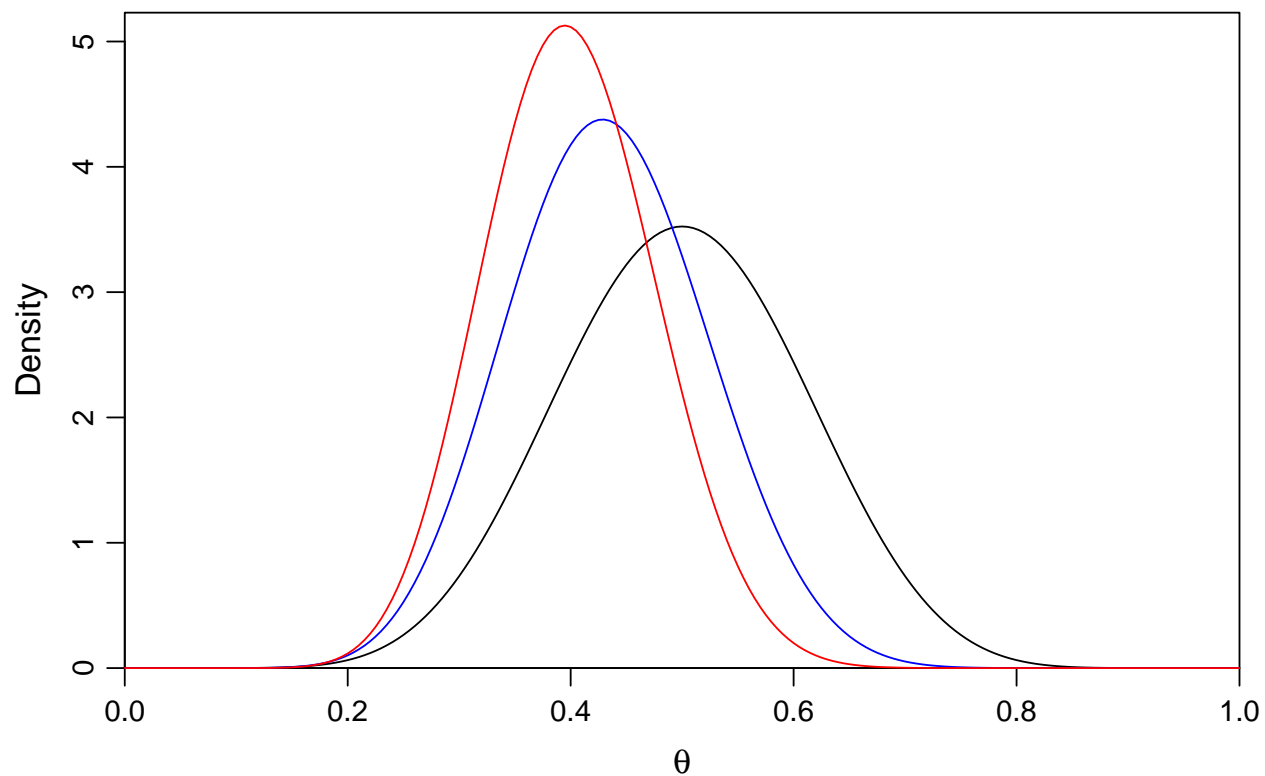


Figure 1: A plot of three beta probability density functions that represent a prior (black;  $\text{beta}(10, 10)$ ), posterior (blue;  $\text{beta}(13, 17)$ ), and empirical Bayes density (red;  $\text{beta}(16, 24)$ ) for a dataset of 10 coin flips, three of which are successes.

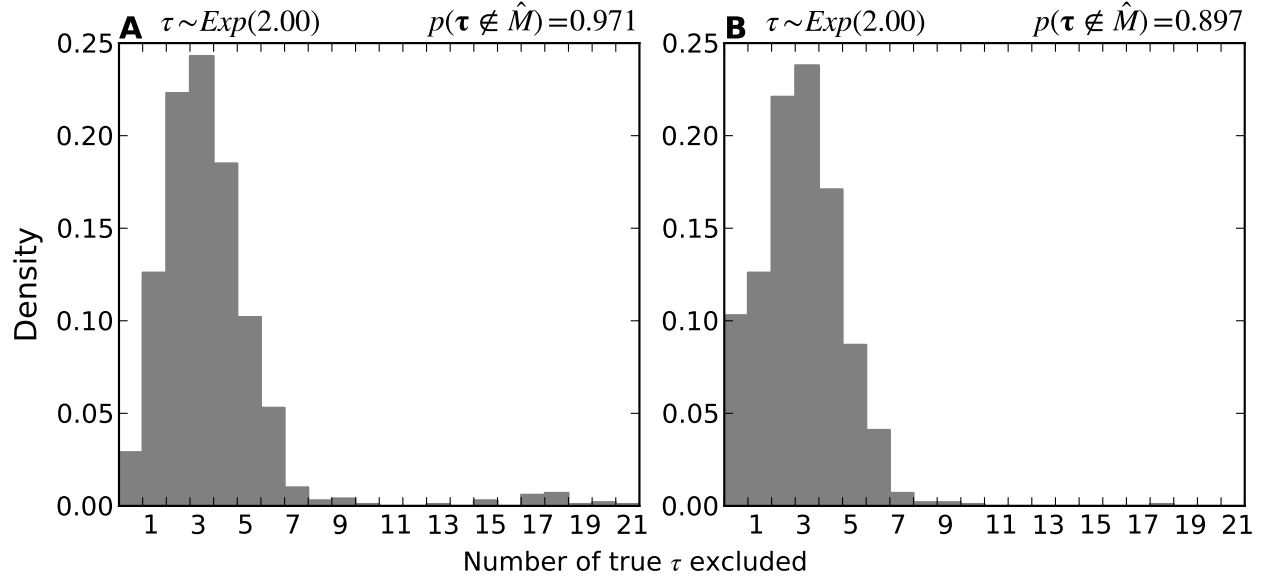


Figure 2: Histograms of the number of true divergence times excluded from the model preferred by the empirically informed model-averaging approach of Hickerson et al. (2014) when applied to simulated datasets in which divergence times of 22 pairs of populations are drawn from an exponential distribution,  $\tau \sim \text{Exp}(2)$ . The plots represent (A) unadjusted and (B) GLM-adjusted estimates from 1000 simulation replicates analyzed using  $5 \times 10^6$  samples from the prior. The proportion of simulation replicates in which at least one true parameter value is excluded from the preferred model ( $p(\tau \notin \hat{M})$ ) is also given.

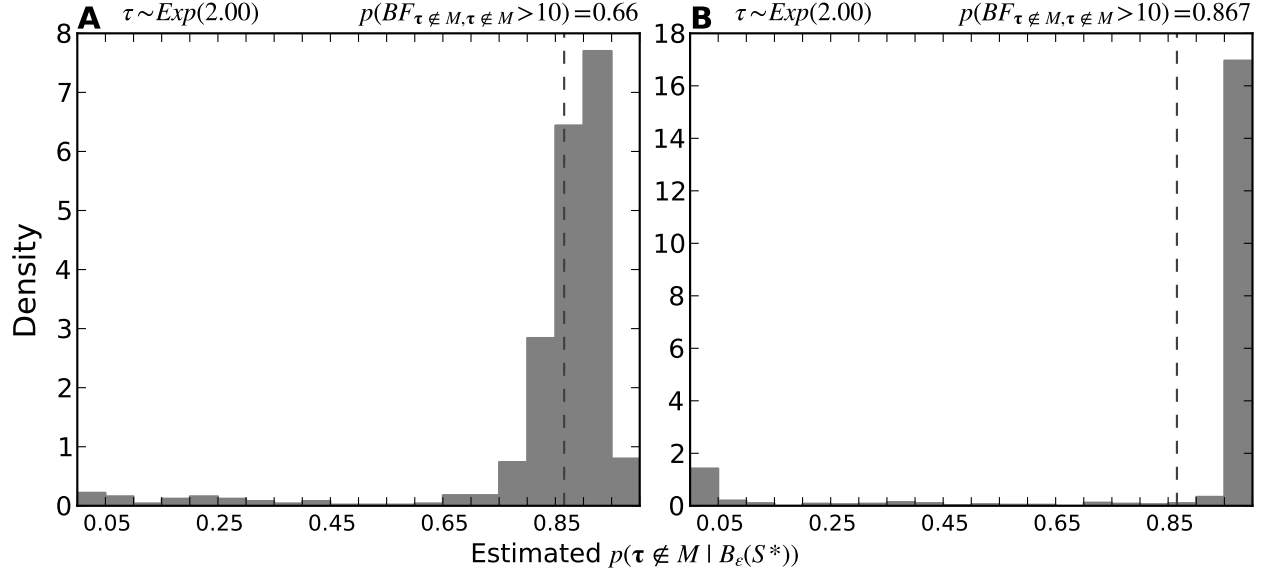


Figure 3: Histograms of the support (estimated posterior probabilities) for excluding at least one true divergence time when the empirically informed model-averaging approach of Hickerson et al. (2014) is applied to simulated datasets in which divergence times of 22 pairs of populations are drawn from an exponential distribution,  $\tau \sim \text{Exp}(2)$ . The plots represent (A) unadjusted and (B) GLM-adjusted estimates from 1000 simulation replicates analyzed using  $5 \times 10^6$  samples from the prior. The proportion of simulation replicates in which there is strong support for at least one true parameter value being excluded from the model ( $p(BF_{\tau \notin M, \tau \in M} > 10)$ ) is also given.

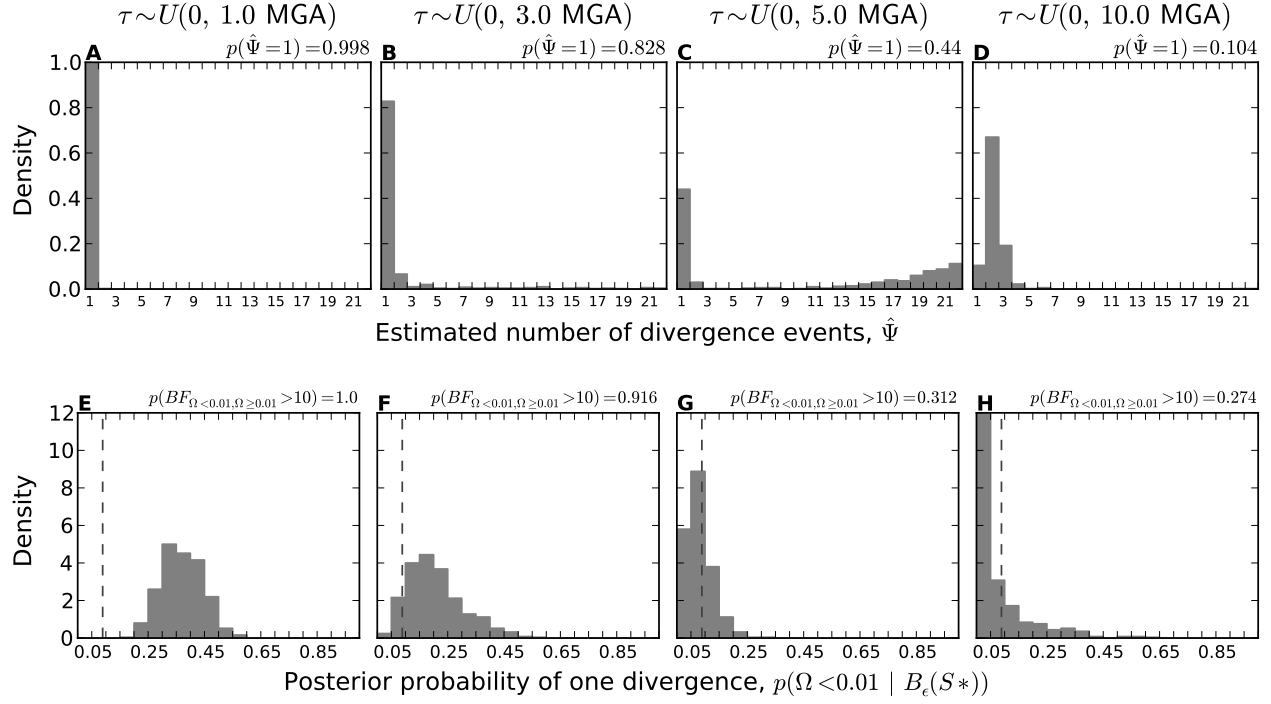


Figure 4: The tendency of the empirically informed model-averaging approach of Hickerson et al. (2014) to (A–D) infer clustered divergences and (E–H) support the extreme model of one divergence when applied to simulated datasets in which the divergence times of 22 pairs of populations are randomly drawn from the uniform distributions  $\tau \sim U(0, \tau_{max})$  indicated at the top of each column of plots (divergence-time distributions are given in units of millions of generations ago (MGA) assuming a per-site rate of  $1 \times 10^{-8}$  mutations per generation). Four of the six  $\tau_{max}$  we simulated are provided; please see Figure S4 for a summary of all of the results.

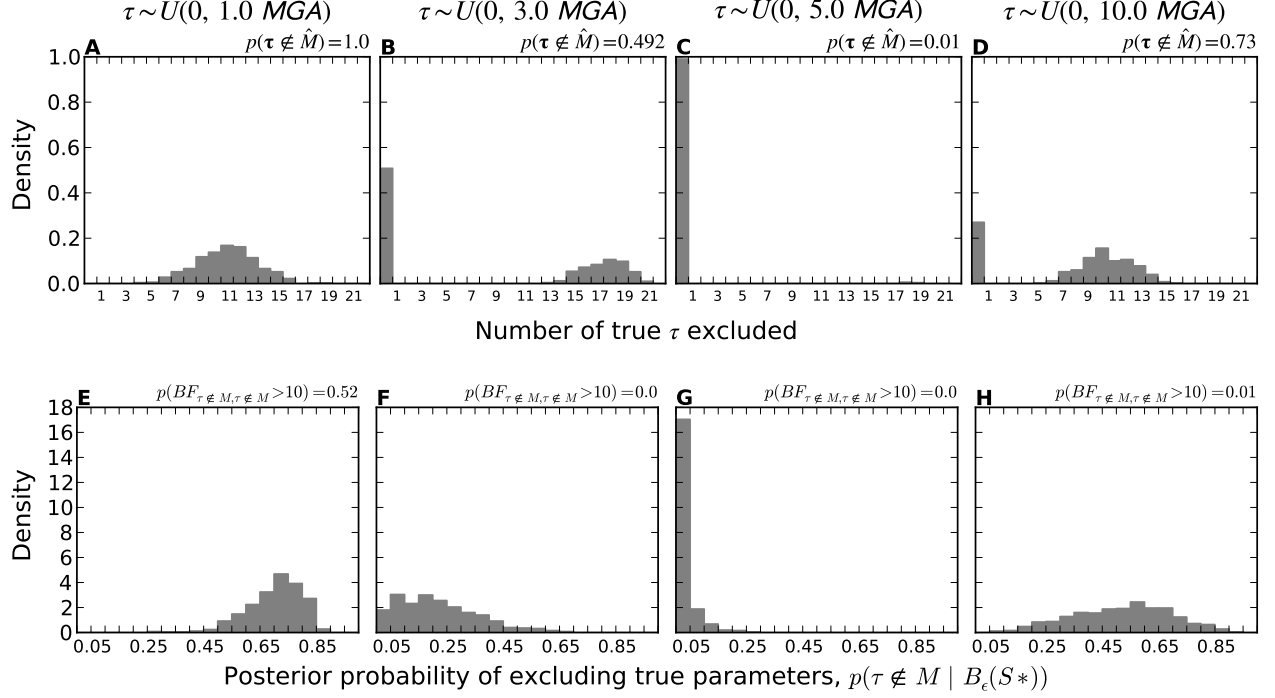


Figure 5: Histograms of the (A–D) number of true divergence-time parameters excluded from the preferred model and the (E–H) posterior probability of excluding at least one divergence-time parameter when the empirically informed model-averaging approach of Hickerson et al. (2014) is applied to simulated datasets in which divergence times of 22 pairs of populations are randomly drawn from the uniform distributions  $\tau \sim U(0, \tau_{max})$  indicated at the top of each column of plots (divergence-time distributions are given in units of millions of generations ago (MGA) assuming a per-site rate of  $1 \times 10^{-8}$  mutations per generation). Four of the six  $\tau_{max}$  we simulated are provided; please see Figure S5 for a summary of all of the results.

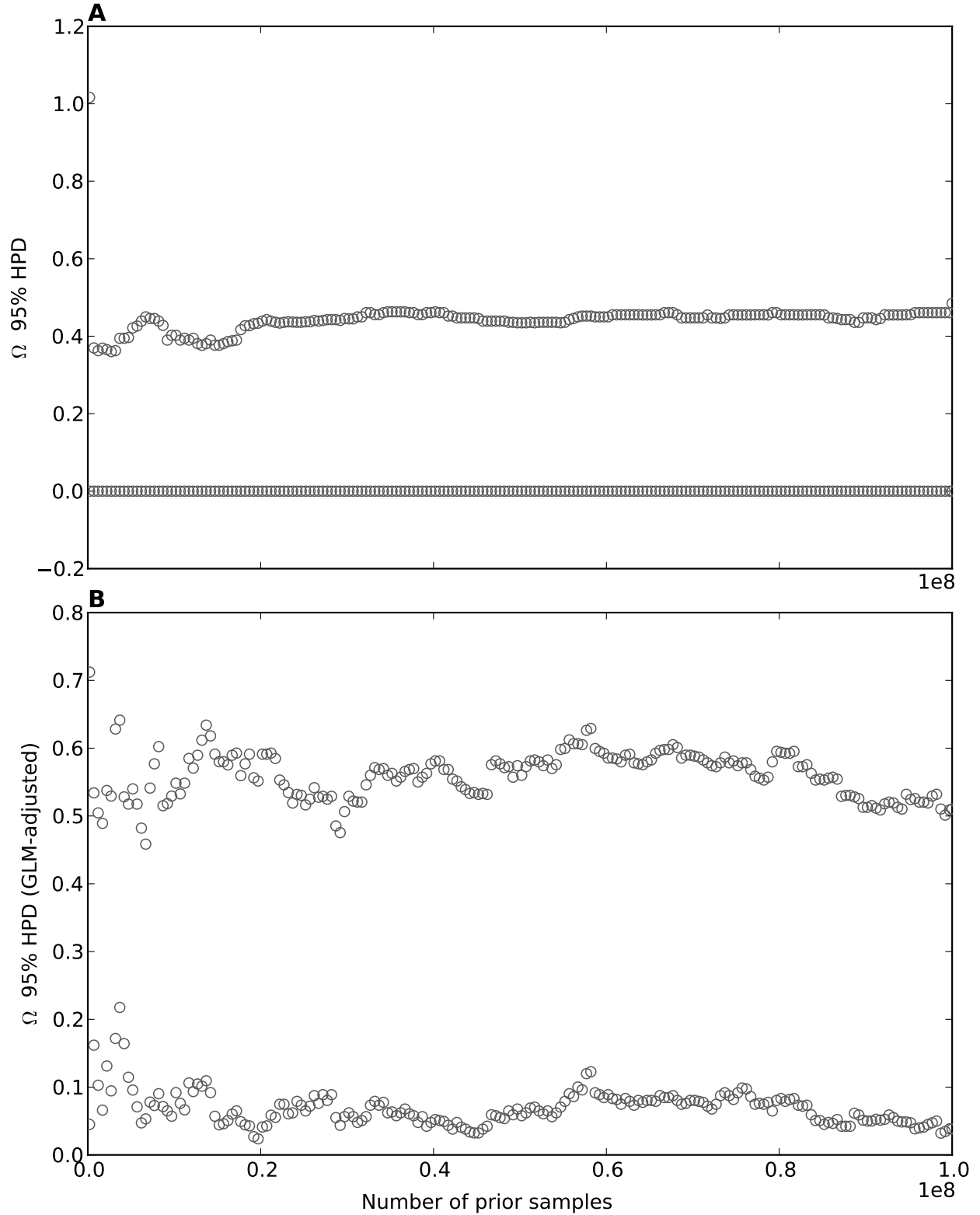


Figure 6: Traces of the estimated lower and upper limits of the 95% highest posterior density (HPD) interval of  $\Omega$  (the dispersion index of divergence times) as 100 million prior samples are accumulated. Each pair of points is based on 1000 posterior samples retained from the prior. Both (A) unadjusted and (B) GLM-regression-adjusted estimates are shown. The data analyzed were the 22 pairs of Philippine taxa from Oaks et al. (2013). Prior settings were  $\tau \sim U(0, 10)$ ,  $\theta_D \sim U(0.0005, 0.04)$ , and  $\theta_A \sim U(0.0005, 0.02)$ .

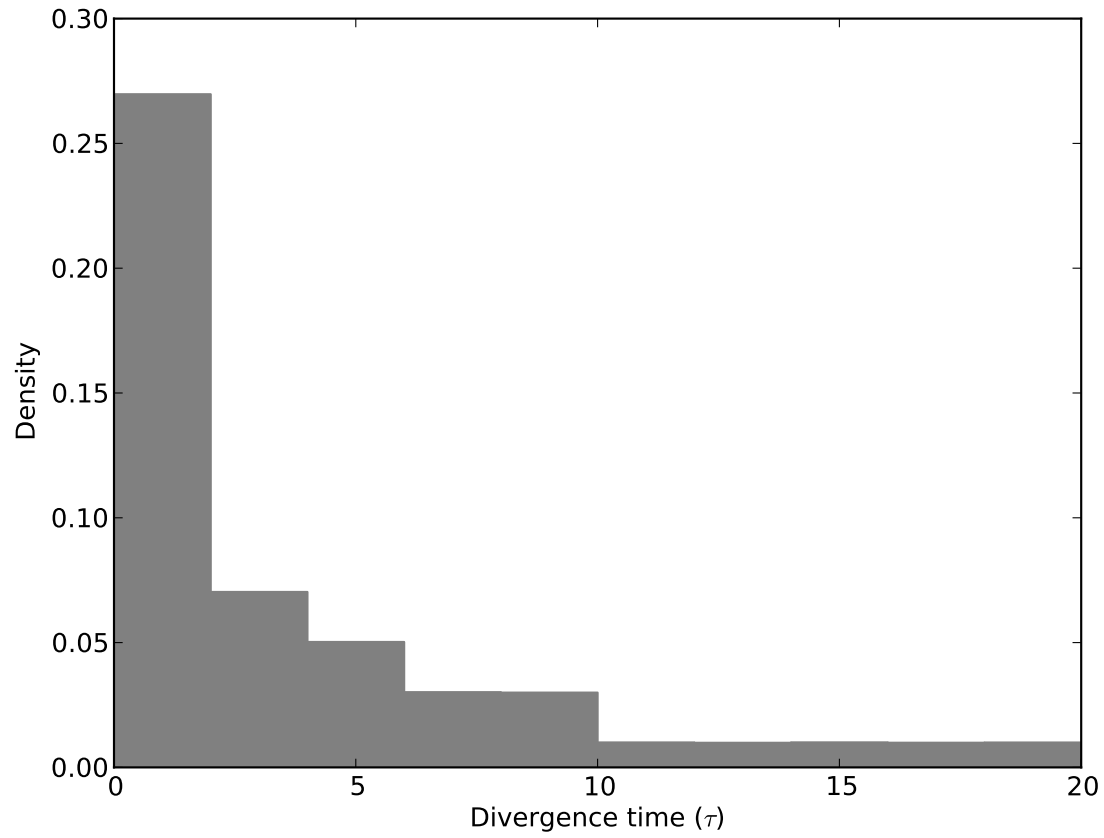


Figure 7: The prior distribution on divergence times imposed by the model-averaging prior comprised of five models with different uniform priors on  $\tau$ :  $M_1$  ( $\tau \sim U(0, 0.1)$ ),  $M_2$  ( $\tau \sim U(0, 1)$ ),  $M_3$  ( $\tau \sim U(0, 5)$ ),  $M_4$  ( $\tau \sim U(0, 10)$ ),  $M_5$  ( $\tau \sim U(0, 20)$ ).



# Supporting Information

Oaks, J. R., C. W. Linkem, and J. Sukumaran. Implications of uniformly distributed, empirically informed priors for phylogeographical model selection: A reply to Hickerson et al.

## 1 An error in Hickerson et al.’s re-analysis of the Philippines data

Hickerson et al. (2014) re-analyzed the dataset of Oaks et al. (2013) using a model-averaging approach, where they placed a discrete uniform prior over eight different prior models (see Table 1 of Hickerson et al. (2014)). However, there was an error in their methodology; their model mixes different units of time.

Each of the eight prior models used in the re-analysis by Hickerson et al. (2014) has one of two priors on the mean size of the descendant populations of each taxon pair:  $\theta_D \sim U(0.0001, 0.1)$  or  $\theta_D \sim U(0.0005, 0.04)$ . As described in Oaks et al. (2013), the divergence-time parameters in the model implemented in **msBayes** are in generations scaled relative to a constant reference-population size,  $\theta_C$ . This reference-population size is defined in terms of the upper limit of the uniform prior on the mean size of the descendant populations,  $\theta_D$ , such that for the prior  $\theta_D \sim U(a_{\theta_D}, b_{\theta_D})$ , the size of the constant reference populations is  $\theta_C = b_{\theta_D}/2$ . Thus, the model used by Hickerson et al. (2014) mixes two different units of time. In other words, some of their prior and posterior samples are in units of  $0.05/\mu$  generations, whereas others are in units of  $0.02/\mu$  generations.

The fact that their posterior samples are in different units makes the results of Hickerson et al. (2014) difficult to interpret, and renders their regression-adjusted results invalid. A fundamental assumption of regression is that all of the values of the response variable are in the same units. Thus, the results in sections “Using ABC Model Comparison to Weight Alternative Priors for the Philippine Vertebrate Data” and “Improved Sampling Efficiency by Prior Weighting Supports Asynchronous and Recent Divergence for the Philippines Vertebrate Data” and presented in Figure 2 of Hickerson et al. (2014) should be disregarded. The error is easily illustrated by re-plotting their results with the different time units indicated (Figure 1).

## 2 Validation analyses

Following Oaks et al. (2013), we characterize the model-choice behavior of the model-averaging approach of Hickerson et al. (2014) under the ideal conditions where the prior is correct (i.e., the data are generated from parameters drawn from the same prior distributions used in the analysis). We used the same prior models as above ( $M_1$ – $M_5$ ; Table 1), and simulated 50,000 datasets under this prior (10,000 from each model). We used a simulated data structure of eight population pairs, with a single 1000 base-pair locus sampled from 10 individuals from each population. We then analyzed each of these replicate datasets using the same prior with 2.5 million samples (500,000 from each of the five prior models), retaining

1000 posterior samples. Our results are very similar to Oaks et al. (2013), but we note that they are not directly comparable as our simulations contained eight population pairs rather than 10 (Figure 6). We find that the approach of Hickerson et al. (2014) estimates the posterior probability of divergence models reasonably well when all assumptions of the method are met (i.e., the prior is correct) and the unadjusted posterior estimates are used. Similar to Oaks et al. (2013), we find that the regression-adjusted estimates of the model probabilities are biased.

### 3 Additional clarifications from Hickerson et al. (2014)

#### 3.1 Saturation of summary statistics

Hickerson et al. (2014) claim the priors used by Oaks et al. (2013) “cause much of the explored parameter space to be beyond the threshold of saturation in most mtDNA genes.” To explore this possibility, we simulated datasets under prior settings that match two of the three priors used by Oaks et al. (2013):  $\theta_D \sim U(0.0005, 0.04)$  and  $\theta_A \sim U(0.0005, 0.02)$ . Under this prior, we randomly sample divergence-time parameters from a uniform distribution of  $U(0, 20)$  coalescent units, simulate datasets, and plot the  $\tau$  values against the summary statistics calculated from the resulting datasets (Figure 7). Clearly, the priors used by Oaks et al. (2013) with upper limits on  $\tau$  of five and 10 coalescent units suffered little to no effect from saturation. Even at divergence times of 20 coalescent units, there is still signal in the summary statistics used by **msBayes** (Figure 7). Thus, the assertion of Hickerson et al. (2014) does not apply to at least two of the priors used by Oaks et al. (2013) and, as a result, does not explain the bias they found.

#### 3.2 Graphical prior comparisons

Hickerson et al. (2014) advocate the use of what they call graphical checks of prior models. This prior-predictive approach entails generating a small number (1000) of random samples from the prior and plotting the resulting summary statistics in comparison to the observed statistics to see if they coincide (see Figure 1 of Hickerson et al. (2014)). Given the richness of the **msBayes** model ( $\approx 600$  parameters for the Philippine dataset analyzed by Hickerson et al. (2014)), we do not expect that 1000 *random* draws from the vast prior parameter space will yield data and summary statistics consistent with the observed data. In fact, when such random draws are tightly clustered around the observed statistics, this can be an indication that the prior is over-fit, as we show in the main text (Table 1 and Figure S2). Thus, using such plots to select priors should be avoided, and the use of posterior-predictive analyses would be much more informative about the overall fit of models.

#### 3.3 Differing utilities of $\Psi$ and $\Omega$ in **msBayes**

The primary component of the **msBayes** model is the vector of divergence times for each of the taxon pairs,  $\boldsymbol{\tau} = \{\tau_1, \dots, \tau_Y\}$  (Oaks et al., 2013). Hickerson et al. (2014) argue that the dispersion index of this vector,  $\Omega$ , is a better model-choice estimator than the number

of divergence-time parameters within the vector,  $\Psi$ . They present a plot of  $\Psi$  against  $\Omega$  (Fig. S1 of Hickerson et al. (2014)), which is essentially a plot of sample size versus variance. This plot shows, not surprisingly, that  $\Omega$  has very little information about the number of divergences among taxa. Nonetheless, Hickerson et al. (2014) conclude  $\Omega$  is more informative and biogeographically relevant than  $\Psi$ . However, all of the information about the temporal distribution of divergences is contained within the divergence-time vector that  $\Omega$  is summarizing. Clearly, the number of divergence time parameters within the vector and their values is more informative than its variance (i.e., the dispersion index is not a sufficient statistic for  $\tau$ ). Hickerson et al. (2014) also argue that “**msBayes** can estimate  $\Omega$  much better than  $\Psi$ .” However, Oaks et al. (2013) demonstrate that even when all assumptions of the model are met,  $\Omega$  is a poor model-choice estimator (see plots B, D & F of Figure 4 in Oaks et al. (2013)), whereas  $\Psi$  performs better.

Importantly,  $\Omega$  is limited to estimating the probability of only a single model (the one-divergence model), and thus its utility for model-choice is very limited. I.e., it can only be informative about the probability of whether there is one divergence shared among the taxa ( $\Omega = 0.0$ ) or there is greater than one divergence ( $\Omega > 0.0$ ). As a result, not only is its model-choice utility limited, but it is also very difficult to estimate.  $\Omega$  can range from zero to infinity, and the point density that it is at its lower limit of zero will always be zero. Thus, an arbitrary threshold (0.01 is used throughout the **msBayes** literature) must be chosen to make the probability of “simultaneous” divergence estimable. Even with this arbitrary threshold, it is still not surprising to see that it is numerically difficult to obtain reliable estimates of the probability that  $\Omega$  is “near” its lower limit of zero. It is easier, less subjective, and more interpretable to estimate the probability of the discrete parameter of the model,  $\Psi$ , is at its lower limit of one. Thus, it is not surprising that Oaks et al. (2013) find that  $\Psi$  is a better estimator of model probability than  $\Omega$ .

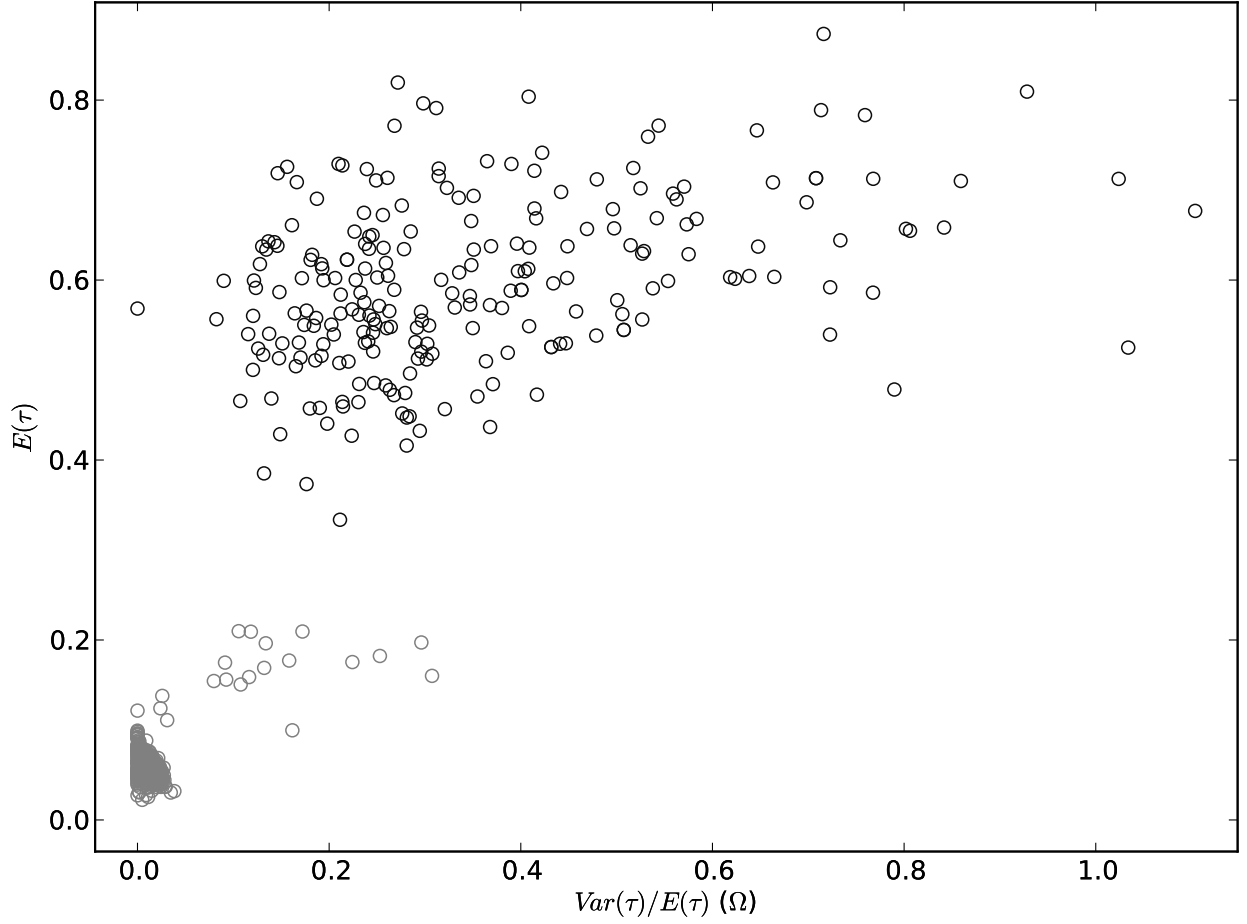


Figure S1: The joint posterior of the mean ( $E(\tau)$ ) and dispersion index ( $\Omega = \text{Var}(\tau)/E(\tau)$ ) of divergence times for 22 vertebrate taxon pairs as estimated by Hickerson et al. (2014) (see Figure 2B of Hickerson et al. (2014)). The posterior samples are color-coded to indicate the erroneous mixture of timescales in the analysis of Hickerson et al. (2014); grey =  $0.05/\mu$  generations and black =  $0.02/\mu$  generations.

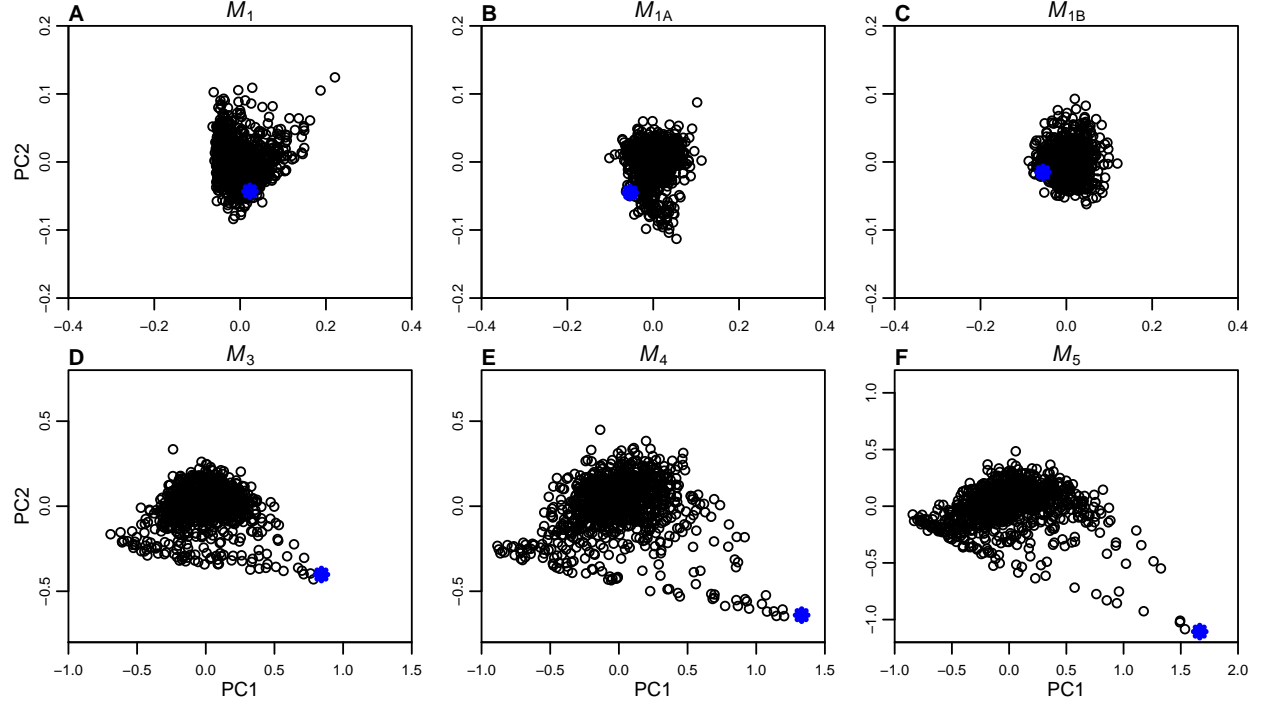


Figure S2: The prior predictive graphical checks recommended by Hickerson et al. (2014) for six prior models: (A)  $M_1$  ( $\tau \sim U(0, 0.1)$ ), (B)  $M_{1A}$  ( $\tau \sim U(0, 0.01)$ ), (C)  $M_{1B}$  ( $\tau \sim U(0, 0.001)$ ), (D)  $M_3$  ( $\tau \sim U(0, 5)$ ), (E)  $M_4$  ( $\tau \sim U(0, 10)$ ), and (F)  $M_5$  ( $\tau \sim U(0, 20)$ ). The three models that likely exclude true values of some divergence times of the 22 pairs of Philippine vertebrate taxa (A–C) appear to have a better “fit” than the priors that likely cover the true divergence times (D–F). The plots project the summary statistics from 1000 random samples from each model onto the first two orthogonal axes of a principle component analysis, with the blue dot representing the observed summary statistics from the 22 population pairs of Philippine vertebrates.

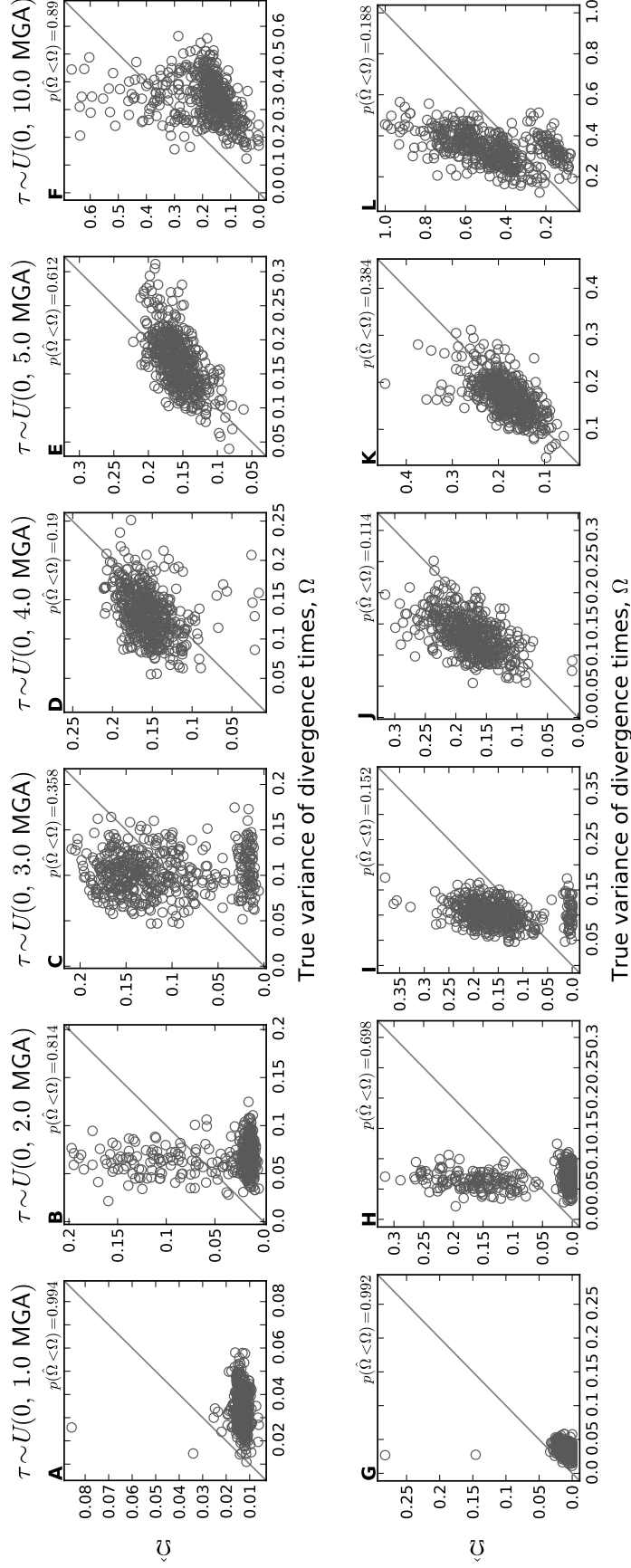


Figure S3: The accuracy of (A–F) unadjusted and (G–L) GLM-adjusted estimates of dispersion index of divergence times ( $\Omega$ ) when the empirically informed model-averaging approach of Hickerson et al. (2014) is applied to simulated datasets in which divergence times of 22 pairs of populations are randomly drawn from the uniform distributions  $\tau \sim U(0, \tau_{max})$  indicated at the top of each column of plots (divergence-time distributions are given in units of millions of generations ago (MGA) assuming a per-site rate of  $1 \times 10^{-8}$  mutations per generation).

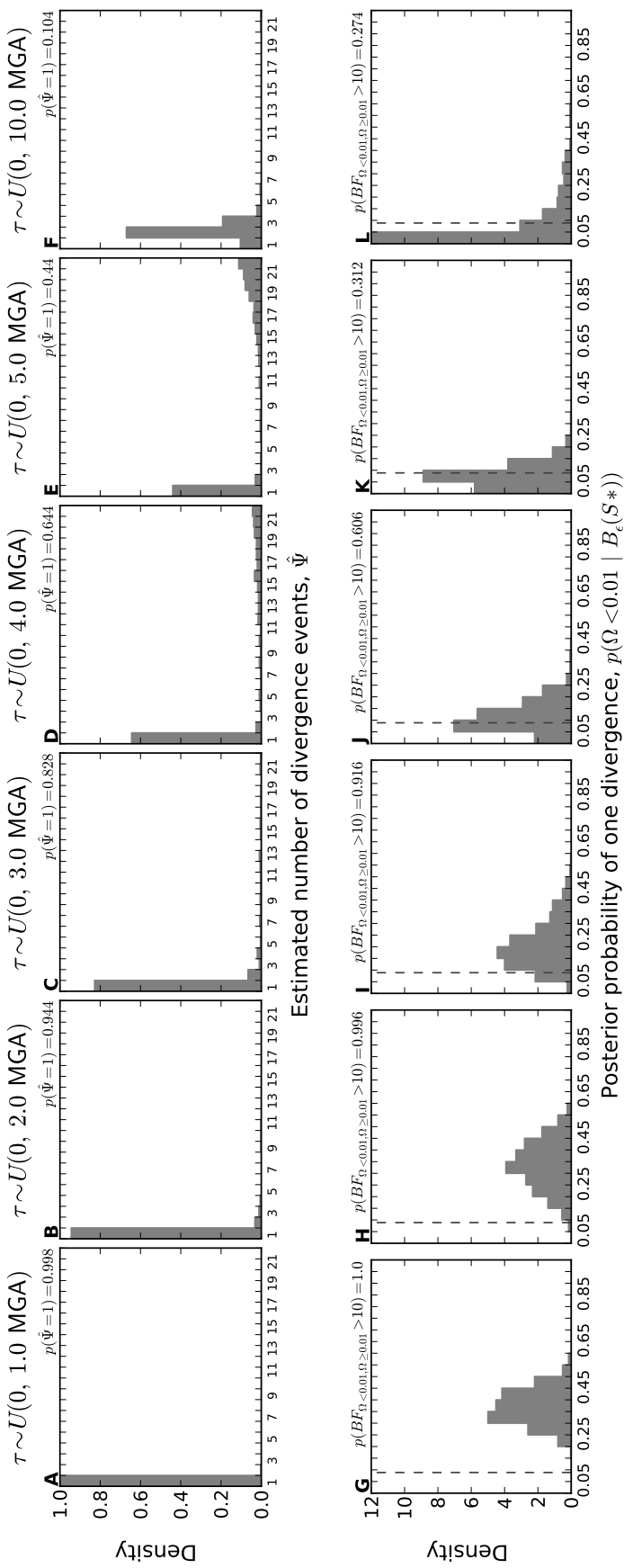


Figure S4: The tendency of the empirically informed model-averaging approach of Hickerson et al. (2014) to (A–F) infer clustered divergences and (G–L) support the extreme model of one divergence when applied to simulated datasets in which the divergence times of 22 pairs of populations are randomly drawn from the uniform distributions  $\tau \sim U(0, \tau_{max})$  indicated at the top of each column of plots (divergence-time distributions are given in units of millions of generations ago (MGA) assuming a per-site rate of  $1 \times 10^{-8}$  mutations per generation).

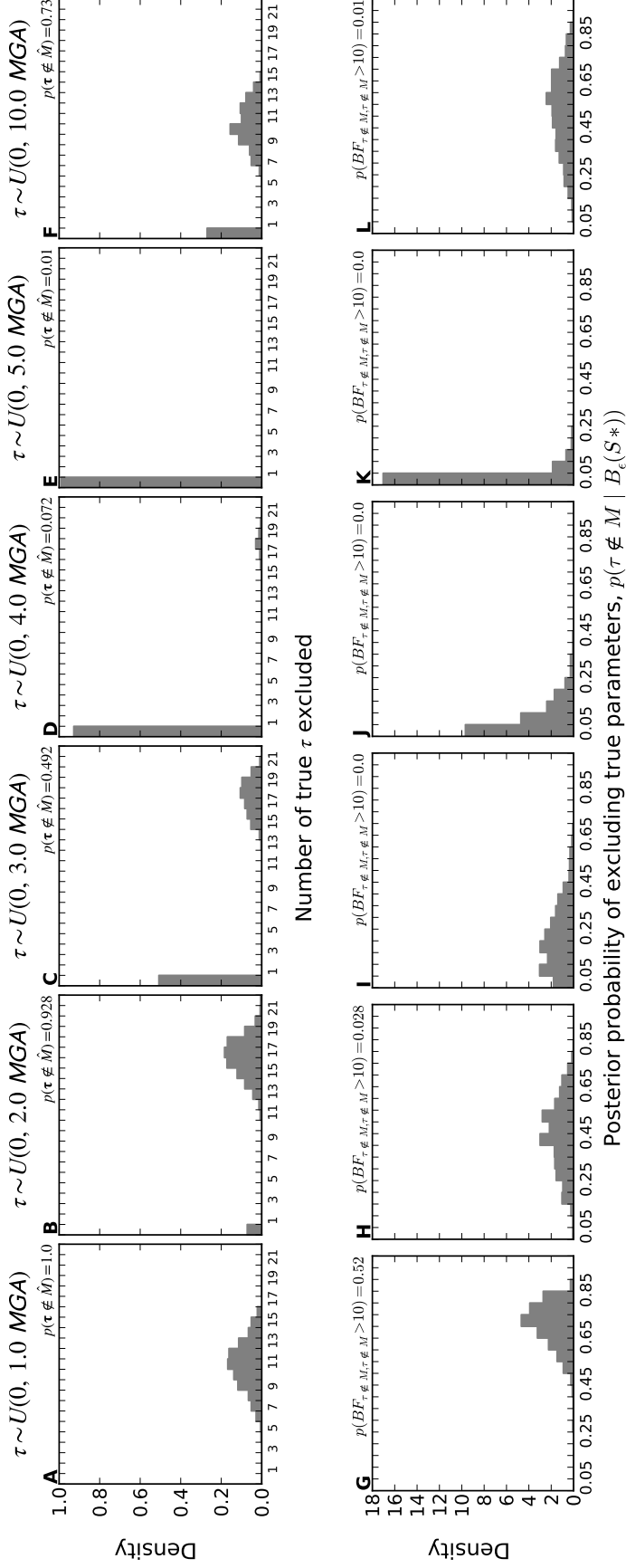


Figure S5: Histograms of the (A–F) number of true divergence-time parameters excluded from the preferred model and the (G–L) posterior probability of excluding at least one divergence-time parameter when the empirically informed model-averaging approach of Hickerson et al. (2014) is applied to simulated datasets in which divergence times of 22 pairs of populations are randomly drawn from the uniform distributions  $\tau \sim U(0, \tau_{max})$  indicated at the top of each column of plots (divergence-time distributions are given in units of millions of generations ago (MGA) assuming a per-site rate of  $1 \times 10^{-8}$  mutations per generation).



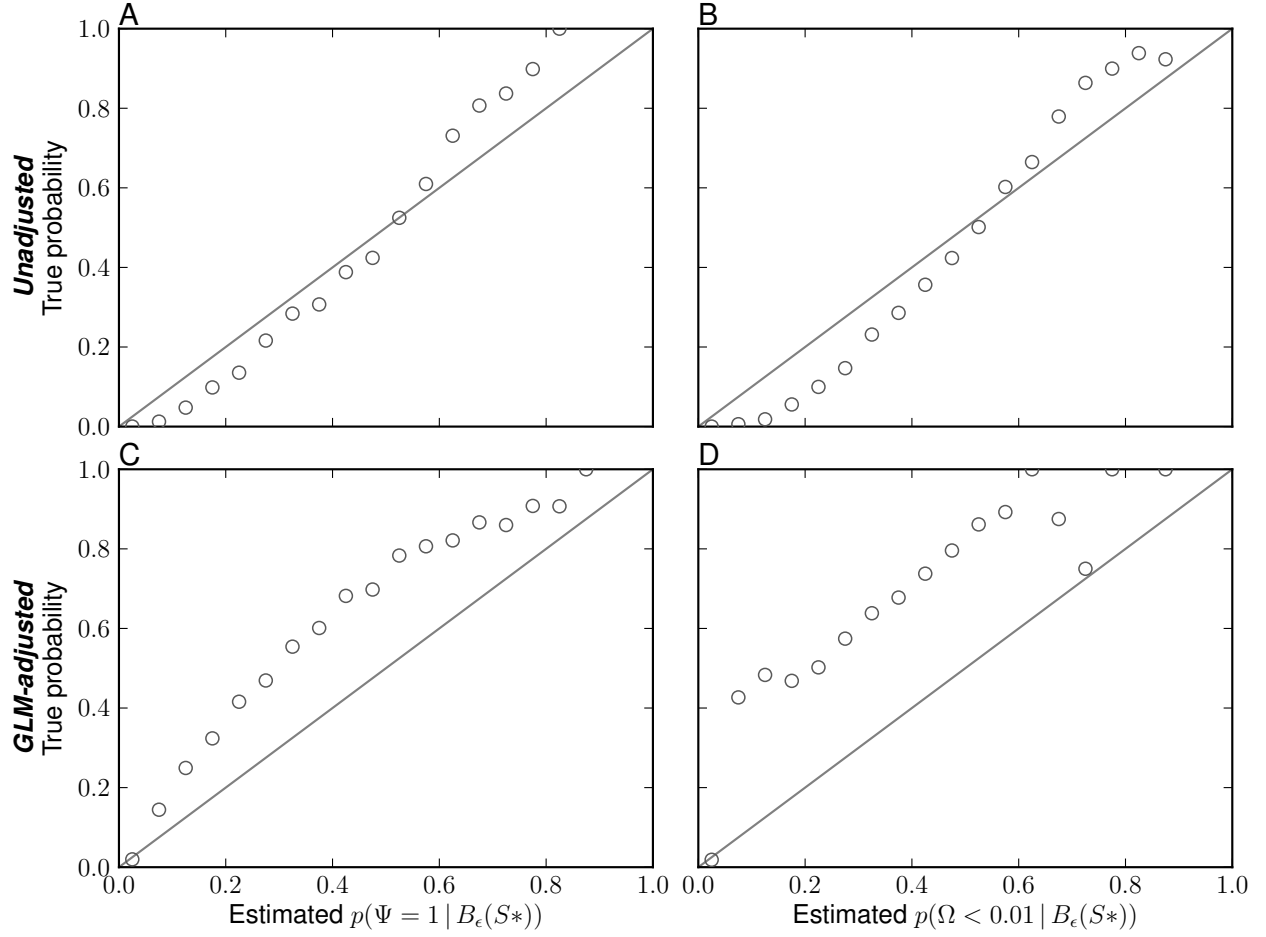


Figure S6: An assessment of the approximate Bayesian model-averaging approach of Hickerson et al. (2014) under the ideal conditions when the prior model is correct (i.e., the datasets are simulated from parameters drawn from the same prior distributions used in the analysis). The plots show the relationship between the estimated posterior and true probability of (A & C)  $\Psi = 1$  and (B & D)  $\Omega < 0.01$ , based on 50,000 simulations. The results summarize the (A & B) unadjusted and (C & D) GLM-adjusted posterior estimate from each simulation replicate. The prior settings for all replicates included five prior models with  $\theta_D \sim U(0.0001, 0.1)$  and  $\theta_A \sim U(0.0001, 0.05)$  for all five models, and  $M_1 : \tau \sim U(0, 0.1)$ ,  $M_2 : \tau \sim U(0, 1)$ ,  $M_3 : \tau \sim U(0, 5)$ ,  $M_4 : \tau \sim U(0, 10)$ , and  $M_5 : \tau \sim U(0, 20)$ . The number of samples from the prior was  $2.5 \times 10^6$ . The simulated data structure was 8 population pairs, with a single 1000 bp locus sampled from 10 individuals from each population. The 50,000 estimates of the posterior probability of one divergence event were assigned to 20 bins of width 0.05. The estimated posterior probability of each bin is plotted against the proportion of replicates in that bin with a true value consistent with one divergence event (i.e.,  $\Psi = 1$  or  $\Omega < 0.01$ ).

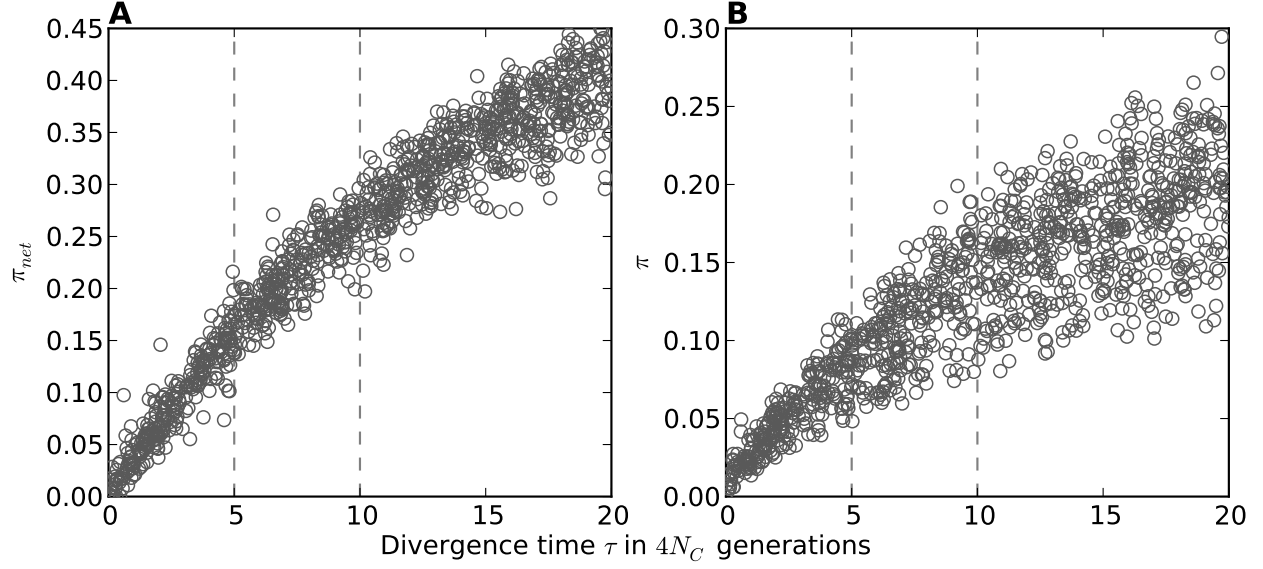


Figure S7: The summary statistics  $\pi$  (Tajima, 1983) and  $\pi_{net}$  (Takahata and Nei, 1985) as a function of divergence time between populations. Each plot represents 1100 pairs of parameter draws and summary statistics calculated from the simulated data. Prior settings for the simulations were  $\tau \sim U(0, 20)$ ,  $\theta_D \sim U(0.0005, 0.04)$ , and  $\theta_A \sim U(0.0005, 0.02)$ .